Classroom Observations from Phase 2 of the Pennsylvania Teacher Evaluation Pilot: Assessing Internal Consistency, Score Variation, and Relationships with Value Added

Final Report

May 31, 2013

Elias Walsh
Stephen Lipscomb

MATHEMATICA
Policy Research

All statistics are calculated by Mathematica unless stated otherwise.

Contract Number:
Agreement dated September 5, 2012

Mathematica Reference Number:
40119.100

Submitted to:
Team Pennsylvania Foundation
100 Pine Street, 9th Floor
Harrisburg, PA 17101
Project Officer: Jennifer Cleghorn

Pennsylvania Department of Education
333 Market Street
Harrisburg, PA 17126
Project Officers: Pat Hardy and Rich Maraschiello

Submitted by:
Mathematica Policy Research
955 Massachusetts Avenue
Suite 801
Cambridge, MA 02139
Telephone: (617) 491-7900
Facsimile: (617) 491-8044
Project Director: Stephen Lipscomb

Classroom Observations from Phase 2 of the Pennsylvania Teacher Evaluation Pilot: Assessing Internal Consistency, Score Variation, and Relationships with Value Added

Final Report

May 31, 2013

Elias Walsh
Stephen Lipscomb

MATHEMATICA
Policy Research

All statistics are calculated by Mathematica unless stated otherwise.

ACKNOWLEDGMENTS

This report would not be possible without the support of the Team Pennsylvania Foundation, the Pennsylvania Department of Education, and the Bill & Melinda Gates Foundation. We are especially grateful to Jennifer Cleghorn, Pat Hardy, and Rich Maraschiello for their leadership throughout this project, and to John Nau, David Ream, and Thresa Stafford for providing us with the statewide data used in this report. We also thank the teachers and principals from the school districts that participated in Phase 2 of the Teacher Evaluation Pilot, because it is through their efforts that we are able to study and improve the teacher evaluation system.

We would like to recognize several Mathematica Policy Research staff who made substantive contributions to this report. Duncan Chaplin, Alexandra Resch, and Brian Gill provided helpful comments and contributions. Clare Wolfendale, Serge Lukashanets, and Elizabeth Keating provided excellent programming support. The paper was edited by Walt Brower and produced by Alexandra Clifford.

Any opinions expressed herein are those of the authors and do not necessarily represent the views of the Team Pennsylvania Foundation or the Pennsylvania Department of Education.

This page has been left blank for double-sided copying.

CONTENTS

TABLES

FIGURES

This page has been left blank for double-sided copying.

EXECUTIVE SUMMARY

A. Rationale for the Study

Like many state education agencies across the country, the Pennsylvania Department of Education (PDE) is engaged in an ambitious effort to revamp methods for evaluating teacher effectiveness. Legislation signed by Pennsylvania's governor in June 2012 (Act 82) mandates that a statewide system of evaluating educators based on multiple measures of performance be developed and implemented starting with the 2013–2014 school year. The law requires that 50 percent of evaluation scores be based on principals' classroom observations of teachers' professional practices. The other 50 percent will be determined by student growth data (15 percent), building-level information (15 percent), and an elective measure (20 percent). The goal of the evaluation framework, as indicated by PDE, is to promote student learning by improving how teacher effectiveness is measured.

This report presents findings from Phase 2 of a three-year teacher evaluation pilot being conducted by PDE. Phase 2 took place during the 2011–2012 school year and included 2,621 teachers from 105 local education agencies whose teaching practices were evaluated by their principals with a rubric that will be used in the new statewide evaluation system. The rubric is based on the Framework for Teaching, which school districts in Pittsburgh, Cincinnati, and other places have adopted for teacher evaluations, and includes 22 components grouped into four broad domains of teaching practice. For Phase 2, teachers were rated on a consistent set of three of these components and on some or all of the remaining components at the principal's discretion. The rubric was selected by a steering committee convened by PDE and Team Pennsylvania Foundation (Team PA) and was piloted initially in just four school districts during 2010–2011 (called *Phase 1*). The rubric is being piloted for a third time in 2012–2013 in about 270 districts (*Phase 3*) ahead of its planned rollout in 2013–2014.

B. Research Questions and Main Findings

For Phase 2, PDE sought to address three research questions using the classroom observation scores from 2011–2012. The questions are aimed at providing information about the rubric's overall usefulness as an evaluation tool, based on its first implementation in a large number of the state's school districts. We describe each research question and summarize key findings below.

1. **Do rubric components contribute to consistent measures of teaching performance and professional practices?** We find that ratings tend to be similar across rubric components in the same domains, and that the consistency of domain-level average scores (regardless of which components are rated) across domains is even better. We do not find any evidence that it is preferable for teachers to be rated on certain components versus others; that is, the consistency of rubric scores is not meaningfully affected by excluding any one component. Our findings support a conclusion that principals' decisions about which components to include in observations may not compromise the fairness of teachers' overall scores. Even so, using more components leads to more consistent scores, so we recommend that principals rate teachers on components whenever evidence exists to support a rating.
2. **Do the rubric ratings of teachers' professional practices vary across teachers so that teachers can be differentiated based on their effectiveness?** We find that, on

most rubric components, the practices of at least 90 percent of teachers in Phase 2 were rated by their principal as either *proficient* or *distinguished*— the highest two of four performance categories. Summary rubric measures that combine scores from multiple components have more potential to differentiate teacher effectiveness because they take on a continuous range of values. However, estimates of teachers' contributions to their students' achievement from student-level data may provide even finer distinctions because measures based on the rubric are likely to be less precise. We also find that most, but not all, of the variation in rubric ratings is among teachers at the same schools rather than between schools, which suggests that principals do differentiate teacher performance using the rubric. Nevertheless, the amount of between-school variation in teachers' classroom observation ratings may be greater than the amount of between-school variation in their contributions to student achievement. One explanation is that principals may be applying the rubric in different ways. Our findings support a conclusion that PDE should continue to support principals through ongoing training on the rubric, but they are not conclusive enough to identify any specific concerns about how the rubric was applied during Phase 2.

3. **Do rubric ratings correspond to teachers' estimated contributions to their students' achievement gains, and if so, are these relationships consistent across grades and subjects?** Across nearly all components, we find that teachers with higher rubric scores tend to make larger contributions to student achievement. The magnitudes of the relationships—correlations of 0 to 0.28—are small, but consistent with those found in previous research on similar rubrics. The largest relationships are for components that measure effective instructional practices; the smallest are for those that measure professional responsibilities. We find positive relationships between rubric scores and value added for both elementary and middle school teachers, with larger-magnitude relationships for middle school science teachers. The relationships for specific grade spans and subjects are based on smaller numbers of teachers, which suggests that the results should be confirmed with study of larger samples. Our findings support a conclusion that the rubric is measuring aspects of teachers' practices that are related to growth in student achievement on standardized assessments. This is consistent with PDE's view that the rubric can be used as a tool to promote student learning.

C. Limitations

Our main findings should be considered along with several limitations of the study.

1. **The sample of teachers in Phase 2 was heavily concentrated in one district.** PDE intended for about 7,000 teachers to participate in Phase 2, but because the time commitment was more than expected, school administrators in the 105 districts ultimately submitted rubrics for only 2,621. The main exception was in Pittsburgh Public Schools (Pittsburgh), which applied the rubric broadly across its teachers, contributing more than 60 percent of the total number of observations collected. Because of the demographics of the collected data, this report presents separate findings for all Phase 2 teachers, for Pittsburgh teachers only, and for non-Pittsburgh teachers only. Since Phase 2 teachers are not a random sample of teachers in the state, our results may not reflect the results we would obtain had all teachers participated. To address this limitation we examine how our results are affected by accounting for observable characteristics of the sample of participating teachers. As best we

can tell, our main findings are not substantially affected by the selection of participating teachers, though we cannot account for all factors related to Phase 2 participation.

2. **Inter-rater reliability and test-retest reliability could not be assessed.** Because teachers in Phase 2 were observed once and by just one rater, we were not able to examine whether independent raters reach similar conclusions about a teacher's professional practices and whether raters apply the rubric consistently across time. PDE includes inter-rater consistency training when principals are initially trained on the rubric, but these two dimensions of the rubric's reliability have such fundamental implications for the fairness of the evaluation system that PDE should collect data on them regularly and in the context of actual observations, especially while the rubric is still new to principals.

3. **Teachers' estimated contributions to their students' achievement gains are based on standardized assessment scores only, and a limited number of grades and subjects.** Our attempt to validate the logic model guiding the rubric is limited by the types of information on student learning that are available on a statewide basis. We used Pennsylvania System of School Assessments (PSSA) scores in all subjects in grades 4 through 8 to construct measures of teacher effectiveness. The findings based on these measures are directly applicable only to teachers in those tested grades and subjects, and do not consider the many other ways in which teachers help students, such as by encouraging them not to drop out.

This page has been left blank for double-sided copying.

I. INTRODUCTION

A. Rationale for the Study

Like many state education agencies across the country, the Pennsylvania Department of Education (PDE) is engaged in an ambitious effort to revamp methods for evaluating teacher effectiveness. Legislation signed by Pennsylvania's governor in June 2012 (Act 82) mandates that a statewide system of evaluating educators based on multiple measures of performance be developed and implemented starting with the 2013–2014 school year. The law requires that 50 percent of teachers' evaluation scores be based on principals' classroom observations of teachers' professional practices. The other 50 percent will be determined by student growth data (15 percent), building-level information (15 percent), and an elective measure (20 percent). The evaluation framework, when completely designed, will combine this information across measures into a final rating score. The goal of the evaluation framework, as indicated by PDE, is to promote student learning by improving how teacher effectiveness is measured.

Prior to the passage of Act 82, PDE and Team Pennsylvania Foundation (Team PA) convened a steering committee in September 2010 to design a rubric—a formal instrument for collecting evidence using a prescribed set of procedures—to evaluate the quality of teachers' professional practices based on principals' classroom observations. The rubric was piloted for the first time during 2010–2011 in four school districts (called *Phase 1*).[1] Following Phase 1, PDE determined that the rubric would be used to rate teachers' professional practices statewide when the new evaluation model is introduced. In 2011–2012 (*Phase 2*), the rubric was piloted in 105 school districts.[2] The rubric is being piloted for a third time in 2012–2013 (*Phase 3*) in about 270 districts ahead of its planned rollout in 2013–2014.

At the request of PDE and Team PA, Mathematica Policy Research has conducted a study of the rubric data obtained during Phase 2 to help guide efforts to refine and enhance this part of the evaluation framework, if needed, before it is implemented across the state.[3] This report describes the findings from our study, which pertain to three main topics:

1. Whether teachers' ratings on rubric components that measure similar constructs reach similar conclusions about performance
2. Whether there is variation in rubric scores to distinguish effectiveness among different teachers
3. Whether rubric ratings correspond to teachers' estimated contributions to their students' achievement gains

[1] Findings from Phase 1 are described in Lane and Horner (2011) and in Lipscomb et al. (2012).

[2] The steering committee, the Phase 1 pilot, and the Phase 2 pilot were supported through grants from the Bill & Melinda Gates Foundation.

[3] The scope of our study is limited to the classroom observation rubric measure of the teacher evaluation framework, though information about the other measures in the framework could also suggest refinements.

The study topics aim to provide an indication of the rubric's overall usefulness as an evaluation tool. By examining the similarity of teachers' scores on components that purport to measure similar constructs, we can establish whether its components are likely to provide a consistent assessment of teacher mastery in certain professional practice areas. By examining the variation in scores, we can describe the extent to which principals' ratings of their teachers' practices based on the rubric tend to be similar or different across teachers. We can also describe whether teachers supervised by some principals appear to be rated higher or lower than teachers supervised by other principals. If the two groups of teachers are equally effective on average, this could signal a need for ongoing training on the rubric. Finally, by studying the correspondence between rubric scores and teachers' estimated contributions to student achievement gains, we can describe the practices exhibited by teachers who are effective at raising their students' standardized achievement scores. We measure teacher effectiveness using value-added models (VAMs). VAMs are statistical models that estimate teachers' contributions to their students' achievement growth, based (in this report) on statewide standardized assessments in math, reading, science, and writing in grades 4 through 8. If teachers with higher rubric scores are also more effective at raising student achievement, it helps to establish the rubric's validity—the extent to which it measures what it intends to measure—because PDE's ultimate goal for the rubric is to promote student learning. A better understanding of these relationships could also help school districts design and target professional development programs to teachers.

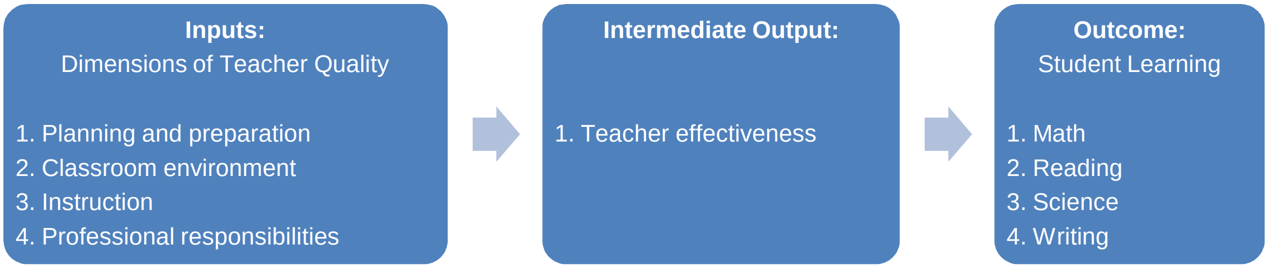
B. The Pennsylvania Teacher Evaluation Pilot, Phase 2

During Phase 2, a sample of Pennsylvania teachers were evaluated using a classroom observation rubric based on the Framework for Teaching, which was developed by Charlotte Danielson and selected for the pilot during Phase 1. The rubric includes 22 components grouped into four domains of teachers' professional practice:

1. **Planning and preparation,** such as effectiveness in demonstrating content knowledge and designing coherent instruction
2. **Classroom environment,** such as effectiveness in creating a positive classroom culture that promotes continuous student growth and development
3. **Instruction,** such as effectiveness in putting into place practices that support student learning
4. **Professional responsibilities,** such as professionalism in conduct and effectiveness in promoting professional growth engaging parents and community members

The domains measure different aspects of teacher effectiveness, which PDE views as an intermediate output in the production of student learning, as indicated by the logic model below:

Figure I.1. Logic Model for Evaluating Teacher Effectiveness Using the Rubric



It is thought that by improving the measurement of the dimensions of teacher quality that are captured by the rubric, schools and districts will be better able to identify and retain effective teachers, promote the development of professional skills associated with student learning, and identify those whose performance is consistently far below proficient levels. This, in turn, will help students to achieve at higher levels.

Teachers and principals in 105 Pennsylvania school districts participated in Phase 2. District participation was voluntary. School principals in participating districts then selected which of their teachers would be observed based on the pilot rubric. PDE advised principals not to include in the Phase 2 pilot teachers whose performance had been rated as unsatisfactory in previous years.

Principals were not required to rate teachers on every rubric component. All teachers were to be evaluated on a consistent set of three components that measured teachers' mastery in planning coherent instruction, engaging students in learning, and using assessment to inform instruction.[4] Principals then were to choose at least five other components such that all teachers were assessed on at least two components from each of the rubric's four domains.[5] On each component that was selected, principals used a four-level scale to rate teachers as failing (0 points), needs improvement (1 point), proficient (2 points), or distinguished (3 points).

Domain-level ratings were not part of the Phase 2 pilot, although principals will assign these scores from the same four-level scale used to rate components when the evaluation system is implemented. The domain scores will be used to calculate a Professional Practice Rating (PPR) for each teacher—the 50 percent of a teacher's final rating score that is determined by the rubric. PDE has not finalized the formula for calculating the PPR, but the current plan is to use a weighted average of domain scores, where domains 1 and 4 will contribute 20 percent weight each and domains 2 and 3 will contribute 30 percent each. PPR scores less than 0.5 would be called failing; scores at least 0.5 but less than 1.5 would be called needs improvement; scores at least 1.5 but less than 2.5 would be called proficient; and scores at least 2.5 would be called distinguished. The PPR score (that is, the numerical score with decimal values) will enter alongside student growth data, building-level information, and the elective measure in calculating teachers' final ratings. We conduct some of our analyses at the component level. For other analyses, we calculate domain-average scores and PPR scores based on those domain averages, even though PDE plans to recommend that domain scores should reflect the preponderance of the evidence in the domain (that is, not necessarily the average component score). Because teachers were not rated on the same set of components in Phase 2, the components we use to calculate these summary scores vary across teachers and may be different than the ones principals would use in actual observations.

Observation data in Phase 2 were obtained on 2,621 teachers from the 105 districts. PDE had originally intended for about 7,000 teachers to participate, but because the rubric took more time to complete than many administrators who had volunteered for the pilot had expected, there were far fewer completed observations than envisioned. (We describe all our study data in Appendix A.)

Pittsburgh Public Schools (Pittsburgh) contributed 64 percent of the total number of observations for Phase 2 (1,673 of the 2,621 records). Typically, principals in Pittsburgh rated

[4] These components were item 1e (planning and preparation domain) and items 3c and 3d (instruction domain).

[5] PDE issued different guidance to principals for Phase 3: principals are supposed to rate teachers using any component for which they feel evidence exists.

teachers using all 22 components as well as two additional components not part of the state rubric. The main exception was for teachers in Pittsburgh's Supported Growth Project (SGP), who were rated on far fewer components. The SGP, which is available to tenured teachers who were evaluated in the prior year, allows teachers to focus their observations on one or more components and retain their other component scores from the previous year. Consistent with the SGP's design, missing component scores for SGP teachers were backfilled with their scores from 2010–2011. Unlike all other Phase 2 districts, principals in Pittsburgh already had one year of experience with the Framework for Teaching rubric, due to the district's own teacher effectiveness initiatives.[6] Other aspects about Pittsburgh's implementation of the rubric were similar to those of other districts, except that Pittsburgh refers to the second-lowest rating as *basic* rather than *needs improvement.*

The Phase 2 sample resembled teachers across Pennsylvania on some characteristics but not others. As indicated in the first two columns of data in Table I.1, Phase 2 teachers are representative of the state in terms of gender and total years of experience. However, Phase 2 teachers were more likely to be African American (and consequently, less likely to be white, Hispanic, or Asian), were less likely to have a master's degree, and tended to earn higher annual salaries.[7] These differences reflect at least partly the large presence of Pittsburgh teachers in the Phase 2 sample, because the same factors describe similarities and differences between Phase 2 teachers in and out of Pittsburgh. That is, Pittsburgh teachers were more likely than other Phase 2 teachers to be African American, were less likely to have a master's degree, and tended to earn higher annual salaries.

Table I.1. Descriptive Statistics on Teachers in Pennsylvania and in the Phase 2 Sample

		Phase 2		
	Pennsylvania	All	Pittsburgh	Not Pittsburgh
Female (%)	72.3	72.7	72.8	72.6
Race/ethnicity				
White (%)	93.2	89.5*	85.0*#	99.1*
African American (%)	5.0	9.3*	13.3*#	0.7*
Hispanic (%)	1.0	0.2*	0.3*#	0.0*
Asian (%)	0.6	0.3*	0.4	0.1*
Other race/ethnicity (%)	0.3	0.6*	0.9*#	0.0*
Total Experience				
Five years or fewer (%)	16.9	17.6	17.3	18.1
More than 5 years (%)	83.1	82.4	82.7	81.9
Educational Attainment				
Master's degree or higher (%)	52.5	38.0*	33.4*#	47.9*
Bachelor's degree (%)	46.5	61.0*	66.3*#	49.5
Less than bachelor's degree (%)	1.0	1.0	0.3*#	2.6*
Annual Salary ($)	63,100	66,600*	71,400*#	56,200*
Number of Teachers	119,989	2,191	1,494	697

Source: Mathematica calculations based on data from Pennsylvania's longitudinal student database.

Notes: Test statistics allow for unequal variances across samples.

* Mean difference relative to all Pennsylvania teachers is statistically significant at the 5 percent level.
# Mean difference between Phase 2 teachers inside and outside Pittsburgh is statistically significant at the 5 percent level.

[6] Pittsburgh's rubric is called the Research-based Inclusive System of Evaluation, which, as described above, includes two district-developed components in addition to the 22 components that are common to the state's rubric.

[7] Information on other teacher characteristics was not available.

C. Research Questions

Using data collected during Phase 2 of the pilot, this study addresses three primary questions:

1. **Do rubric components contribute to consistent measures of teaching performance and professional practices?** We measure the extent to which teachers receive similar ratings on components measuring similar concepts (that is, those in the same domain). If all components contribute to consistent measures, then the differences in the components that make up teachers' overall scores might not substantially impact the fairness of the rubric.
2. **Do the rubric ratings of teachers' professional practices vary across teachers so that teachers can be differentiated based on their effectiveness?** Good observational measures differentiate high and low performance. If teachers' scores are all very similar, or if principals implement the rubric differently, then the rubric cannot distinguish performance between teachers well.
3. **Do rubric ratings correspond to teachers' estimated contributions to their students' achievement gains, and if so, are these relationships consistent across grades and subjects?** If the rubric accurately measures teacher practices that are related to student outcomes, then higher rubric scores should relate positively to other indicators of teaching effectiveness based on student outcomes. A consistent set of relationships across grades and subjects supports a conclusion that the rubric can be applied broadly across teachers who serve different groups of students.

D. Related Literature

1. Reliability and Internal Consistency of Subjective Performance Measures

As school systems adopt teacher evaluation systems that rely on new measures of effective teaching, the reliability of these measures has become an important research topic. Reliability describes how well differences between teachers' scores on a measure reflect differences in scores observed at a later time, by a different rater, or on different components of the measure. For example, test-retest reliability is calculated by comparing the same teachers' scores as measured by the same rater, but based on observing different lessons. Another kind of reliability—*inter-rater*—is calculated on the basis of different raters of the same teacher. Whereas both these types of reliability depend on multiple observations of the same teacher, *internal consistency* (a third type of reliability, which measures the similarity of scores across components of a performance measure) is calculated on the basis of a single observation for each teacher. These measures of reliability address in different ways the proportion of dispersion in teachers' scores that can be attributed to persistent differences in their practices rather than differences due only to idiosyncrasies (that is, of the observed lesson, rater, or components of the measure, respectively). No single measure of reliability can address the full range of these concepts on its own.

Prior research on the reliability of classroom observation scales for evaluating teacher performance has focused on measuring test-retest reliability and inter-rater reliability. The Measures of Effective Teaching (MET) project combines these two types of reliability by measuring, in teachers' scores, the percentage of dispersion that could be attributed to persistent differences in teacher practice rather than differences due only to idiosyncrasies of the rater or the observed lesson. Based on this measure, they find that the reliability of the Framework for Teaching—the subjective performance evaluation rubric on which the Phase 2 rubric is based—is higher when a teacher is evaluated by multiple raters and/or on multiple lessons (Kane and Staiger 2012; Ho and Kane 2013). For example, the proportion of dispersion in teachers' scores attributable to persistent differences in teacher practice rose from 0.43 to 0.57 when teachers were rated by two raters based on two different lessons instead of only a single rater and lesson.[8]

Our contribution to prior work on the reliability of teacher performance measures is to calculate the internal consistency of the Phase 2 rubric. We identify three reasons to focus on internal consistency. First, prior work evaluating the reliability of teacher performance measures has typically not examined internal consistency, so this analysis provides new information about how classroom observation measures perform on this dimension of reliability.[9] Second, because Phase 2 teachers were evaluated only once, we cannot calculate reliability measures that require multiple observations of the same teachers. Third, because principals did not need to use all the components in each domain during Phase 2, it is critical that components that are rated contribute to a consistent overall evaluation score. If the rubric does not show a high degree of internal consistency, this could limit other dimensions of reliability for the rubric and raise concerns about its fairness.

Understanding how individual components affect the internal consistency of the rubric will help PDE to assess whether components that are grouped together into domains measure similar constructs. The analysis also could inform observation measures based on the Framework for Teaching in other states or districts. Our focus on consistency of the rubric rather than other measures of reliability does limit our conclusions about the reliability of the rubric. Internal consistency cannot be compared to inter-rater or test-retest reliability, and does not provide information about how much a teacher's score would change if the teacher were rated again using the same components. Internal consistency does describe how similar a teacher's score is likely to have been had it been based on a different set of components.[10]

2. Variation in Performance Measures Across Teachers

Recent work also informs our second research question, which is whether the Phase 2 rubric can differentiate teachers based on the quality of their teaching practices. The New Teacher Project (Weisberg et al. 2009) found that typical subjective performance ratings produced almost universally positive evaluations of teachers. In some cases, even new teacher evaluation systems have not

[8] Sturman et al. (2005) conducted a meta-analysis of research published prior to 2003 on the reliability of subjective and objective job performance measures in several occupations. While none of the studies included teachers, they find reliabilities as high as 0.78. Reliabilities were smaller in occupations rated as more complex.

[9] We are aware of only one prior analysis of the internal consistency of a rubric based on the Framework for Teaching. Benjamin (2002) found internal consistencies between 0.84 and 0.93 for components in each of the four Framework domains. In addition, another Mathematica team, led by Duncan Chaplin, is examining the internal consistency of rubric ratings in Pittsburgh for the U.S. Department of Education's Mid-Atlantic Regional Education Laboratory. That study, when completed, will also include correlations between teacher effectiveness data based on student surveys and both rubric scores and value-added estimates.

[10] This interpretation of internal consistency is based on the assumption that the components each teacher is rated on were chosen randomly. For example, a principal may nonrandomly select components—intentionally or not—on which a teacher has similar skills, leaving out components that if selected would lead to less consistency across components. In future work, it might be possible to determine how component selection affects internal consistency by comparing differences in ratings on one component between teachers who are and are not rated on another component. We could not conduct this analysis in Phase 2 because a large number of teachers did not participate, and therefore it is difficult to distinguish teacher selection from component selection among participating teachers.

resulted in a wider distribution of teacher evaluation scores (Sawchuk 2013). If evaluations do not reasonably differentiate teachers, they cannot inform human resource management decisions. In contrast to the type of evaluations examined by the New Teacher Project, more comprehensive evaluation tools that identify and evaluate specific teaching practices and provide multiple rating categories do allow principals to distinguish between the most- and least-effective teachers in their schools (Jacob and Lefgren 2008; Harris and Sass 2010; Rockoff and Speroni 2011). The MET project found that observers assigned few teachers scores in the lowest or highest performance categories of the Framework for Teaching, but on some components over half of teachers were rated in the lowest two categories combined (Kane and Staiger 2012). Lipscomb et al. (2012) found less variation in rubric scores among the 153 teachers who were part of Phase 1 of the Pennsylvania Teacher Evaluation Pilot, where 96 percent were rated as either *proficient* or *distinguished*.

We extend this work in three ways. First, we describe the variation in scores from a larger implementation of the Framework for Teaching in Pennsylvania than was done in Phase 1. This broader sample allows us to compare variation in scores in Pittsburgh—where the rubric already had been implemented for all teachers in the district, and where principals may be more familiar with applying it—to variation in other Pennsylvania districts. Second, we compare the variation in rubric scores to variation in value-added measures of teacher effectiveness. VAMs are designed to isolate the effect of a teacher's teaching from other factors affecting student achievement including the background characteristics of the students (Meyer 1997; Sanders 2000; McCaffrey et al. 2004; Raudenbush 2004; Hanushek et al. 2005; Kane and Staiger 2008; Glazerman et al. 2010). Value added has been shown to meaningfully differentiate teachers based on their contributions to academic achievement (Rockoff 2004; Hanushek et al. 2005; Aaronson et al. 2007; Chetty et al. 2011). Third, we investigate the extent to which principals' ratings in Phase 2 varied within versus between schools. If the amount of between-school variation is larger than expected—for example, as compared to the same amount for value added—this could indicate that principals do not apply the rubric consistently. The results could suggest whether PDE should consider providing additional training for evaluators.

3. Validity of Performance Measures

Researchers have examined the validity of observational measures similar to the Phase 2 rubric. Validity pertains to how well differences in scores across teachers on one measure reflect differences in scores on a second measure that is assumed to be the key quantity of interest. Because the ultimate goal of the evaluation system is to improve student outcomes, we use teachers' contributions to their students' achievement gains as the benchmark for assessing validity. Several studies have compared teachers' scores on observational measures to teacher effectiveness estimates based on a VAM. The MET project finds correlations under 0.2 between Framework for Teaching scores and value added in math and English language arts (ELA) (Kane and Staiger 2012). The authors conclude that these correlations are too low to support using the observations to identify teachers who are effective at raising student achievement. Kane et al. (2011) find similar results for the rubric that is used to evaluate teachers in Cincinnati, which is also based on the Framework for Teaching. Milanowski (2004) and Kimball et al. (2004) find correlations as high as 0.4 between rubrics based on the Framework for Teaching and value added. Similar correlations with value added also have been found for other subjective evaluations by principals or mentors (Harris and Sass 2010; Rockoff and Speroni 2010; Grossman et al. 2010). For instance, Jacob and Lefgren (2008) found correlations of 0.3 between value added and principals' evaluations of teachers.

The analysis of the Phase 2 rubric contributes in three ways to the literature on the validity of subjective evaluations of teachers. First, we extend the analysis to a new setting, Pennsylvania. To the extent that the Phase 2 rubric is similar to the rubrics studied by Kane et al. (2011) and Kane and Staiger (2012), we should expect to find similar results. However, other factors related to the pilot's implementation—including the quality of observer training for principals, and principals' latitude to select rubric components—could affect the validity of the rubric. Second, our results may identify which components of the rubric are most strongly related to value added. Such components could represent particularly promising practices for professional development of teachers and important components of a teacher evaluation system.[11] Finally, we analyze the validity of the Phase 2 rubric separately for teachers in elementary and middle school, and by subject for departmentalized staff. Certain professional practices may be more related to value added in some grades or subjects. If so, then identifying these practices could be useful for targeting professional development strategies.

[11] Kane et al. (2011) find some evidence that components in a domain similar to the Phase 2 classroom environment domain are more related to teachers' contributions in math than are components in other domains studied.

II. THE INTERNAL CONSISTENCY OF PROFESSIONAL PRACTICE SCORES

Internal consistency, or the degree to which a group of components are related, is the single dimension of reliability that we can consider in this report. As mentioned earlier, we would need ratings from multiple raters to determine inter-rater reliability; we would need multiple ratings on the same teacher to determine test-retest reliability. The standard way to measure internal consistency is to calculate Cronbach's alpha,[12] which indicates the degree to which it is permissible to replace a group of components with a summary measure. Cronbach's alpha ranges between zero and one, where larger values are associated with higher levels of internal consistency (Cronbach 1951).

Researchers have reached different conclusions about what value of alpha should be considered "good." The answer can depend on a scale's use. For instance, Wasserman and Bracken (2003) recommend that alpha values should exceed 0.8 for psychological assessment scales if the scales have high-stakes consequences for individuals. Many other researchers have recommended 0.7 as an acceptable alpha value (Cortina 1993). Since we are not aware of any existing guidelines for the internal consistency of teacher evaluation measures, we apply David de Vaus's recommendation, in his widely cited textbook on surveys in social research, that alpha values above 0.8 are considered good, and alpha values above 0.7 are considered acceptable (de Vaus 2002).

A. Phase 2 Rubric Summary Measures Have Acceptable or Good Levels of Internal Consistency

Based on de Vaus's guidelines, summary measures based on the Phase 2 rubric have acceptable or good levels of internal consistency. In Table II.1, we present Cronbach's alpha for each domain in the rubric and for the PPR overall. Separate statistics are presented for all teachers in Phase 2, for the Pittsburgh teachers only, and for the non-Pittsburgh teachers only, along with the number of teachers with an observation contributing to each statistic. The top four rows indicate that rubric domains have an acceptable level of internal consistency across samples. The bottom row indicates that teachers' overall PPRs (that is, the weighted averages of their domain scores) have a good level of internal consistency.

An acceptable level of internal consistency within each domain is important because not all teachers are rated on all components. Cronbach's alpha typically relies on having a sample of individuals with complete data across the components being examined. As a result, the teacher samples for the domain-level internal consistency estimates include only those teachers who were rated on all the components in a given domain. However, few teachers in Phase 2 were evaluated on exactly the same sets of components. Although principals' selection of components precludes a consistent sample for interpretation, this type of selection may describe the way in which the rubric is used when the evaluation model is introduced formally. That is, PDE plans to recommend to principals that they should select only the components for which evidence of the effectiveness of a teacher's practice exists. For the sake of fairness, it would be undesirable if teachers' evaluation scores were influenced by which specific components are rated. The acceptable level of internal

[12] The formula to calculate Cronbach's alpha for a weighted average of components is $\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K} w_{c_i}^2 \sigma_{c_i}^2}{\sigma_S^2}\right)$, where *K* is the number of components, $\sigma_S^2$is the total variance of the weighted average, w_{c_i} is the weight on component *i*, and $\sigma_{c_i}^2$is the variance of component *i*.

consistency within each domain is one indication that the selection of components may not substantially compromise the fairness of the rubric, though this conclusion could be stronger had principals been randomly assigned to rate teachers on specific sets of components.

Table II.1. Internal Consistency of Rubric Domains and of the Professional Practice Rating

	Number of Items in Scale	All Phase 2		Pittsburgh		Not Pittsburgh	
		I.C.	Obs.	I.C.	Obs.	I.C.	Obs.
Domain 1: Planning and Preparation	6	0.78	1,659	0.76	1,063	0.77	596
Domain 2: Classroom Environment	5	0.75	1,639	0.75	1,054	0.73	585
Domain 3: Instruction	5	0.72	1,646	0.68	1,058	0.71	588
Domain 4: Professional Responsibilities	6	0.75	1,440	0.71	938	0.76	502
Professional Practice Rating (PPR)	4	0.84	2,487	0.81	1,572	0.86	915

Source: Mathematica calculations based on Phase 2 classroom observation data from 2011–2012.

Notes: Internal consistency is measured by Cronbach's alpha.

Internal consistencies for domain-level scores are based on teachers with observation data for all rubric components in the domain.

Internal consistencies for the PPR are based on teachers with observation data on at least one component in each domain. Domain-level scores for domains 1 and 4 are weighted 20 percent each, and domain-level scores for domains 2 and 3 are weighted 30 percent each.

I.C. = internal consistency; Obs. = number of observations (teachers).

To make the most of the available data, for the PPR we calculated Cronbach's alpha by treating domain-level scores as the item scores, regardless of which components were used to generate them for each teacher. This approach results in fewer missing data, because most teachers have a score on at least one component from every domain. The approach is also consistent with PDE's current plans to calculate the PPR as a weighted average of domain scores.

The internal consistency for the PPR is higher than for any of the individual domains. The alpha value of 0.84 reported in Table II.1 is based on fewer items than any of the domain-level alphas (4 domains versus 5-6 components per domain), and Cronbach's alpha is systematically higher for scales that include more items.[13] This systematic positive relationship arises because scales become more reliable when they include more components measuring similar constructs. That is, a scale with three components provides a more reliable summary indicator of a construct than can a scale with two components, provided that all the components pertain to the construct of interest. When viewing internal consistency just as a measure of the similarity of components and not so much as a form of reliability, an increase due to adding components to the scale can be misleading. In the case of the PPR, its internal consistency score is higher than any of the domain-level alphas despite having fewer components in the scale. Overall, the findings in Table II.1 support PDE's current plans to calculate the PPR using domain scores, even if domain scores are obtained based on different numbers or sets of components across teachers. That said, increasing the number of components that go into teachers' PPR scores leads to higher internal consistency ratings, so our findings are not meant to suggest that principals should minimize the number of components that make up their classroom observations.

[13] For instance, Cronbach's alpha is 0.91 if all 22 components are grouped into the same scale among the 1,292 teachers who were rated on all components.

B. Specific Rubric Components Do Not Contribute Disproportionately to Internal Consistency

We supplement the main findings on internal consistency by considering how sensitive rubric domains and the PPR are to the exclusion of certain components. Calculating these "leave-out alphas" provides another indication about whether test scales become appreciably different when teachers are rated on different sets of components. We expect that leave-out alphas will be somewhat smaller than the alphas reported in Table II.1, because each of the scales includes one fewer component. As described above, alpha typically increases with the number of items. But if a specific leave-out alpha is notably smaller than the alphas based on the full scales, it would be an indication that the left-out component is very important to measuring the construct of interest. Similarly a notably larger leave-out alpha would indicate that the excluded component is not important to measuring the construct of interest.

The findings, depicted in Table II.2 for Pittsburgh and non-Pittsburgh teachers in Phase 2, suggest that all components and domains may be useful in describing their constructs of interest. The first section of the table shows the leave-out alpha values, where the excluded component is indicated by the column headings. The resulting internal consistency values are consistently slightly lower than the alphas in Table II.1, and in the case of the third domain, below 0.7. However, the main finding is that no single component exerts an overwhelming positive or negative influence over the internal consistency rating of domains in the rubric. The findings are similar in the second section of Table II.2, where domains are excluded in turn from internal consistency calculations for the PPR.

Table II.2. Leave-Out Alphas for Measuring the Internal Consistency of Rubric Domains and the Professional Practice Rating

	Excluded Component					
	A	B	C	D	E	F
Domain 1: Planning and Preparation	0.74	0.76	0.74	0.76	0.73	0.75
Domain 2: Classroom Environment	0.70	0.71	0.70	0.69	0.75	n/a
Domain 3: Instruction	0.68	0.66	0.64	0.68	0.68	n/a
Domain 4: Professional Responsibilities	0.71	0.74	0.75	0.70	0.70	0.71
		Excluded Domain				
		1	2	3	4	
Professional Practice Rating		0.79	0.80	0.77	0.82	

Source: Mathematica calculations based on Phase 2 classroom observation data from 2011–2012.

Overall, we find from Table II.1 that the rubric domains have at least acceptable levels of internal consistency. The internal consistency of entire rubric domain averages appears to be even better, and no single component or domain appears to contribute disproportionately to a smaller or larger level of consistency. This is reassuring given that the rubric is based on the Framework for Teaching, which has been adapted for teacher evaluation systems by other districts and states. While our results are encouraging, we caution readers not to generalize the findings to other forms of reliability that we could not examine. In particular, given prior evidence from other states that the reliability of similar rubrics improves with additional observations and raters (Kane and Staiger 2012; Ho and Kane 2013), it would be very useful to assess inter-rater and test-retest reliability in the future on actual observations, to ensure that the rubric is being applied consistently across raters and time.

This page has been left blank for double-sided copying.

III. THE EXTENT OF VARIATION IN RUBRIC SCORES

Score variation, or the extent to which principals use different rubric ratings to describe differences across teachers in their professional practices, is a prerequisite for an evaluation system to be able to distinguish the degrees of effectiveness among teachers. If rubric scores show little variation across teachers who vary in their effectiveness, then the PPR will have limited usefulness in explaining differences in teachers' final ratings when combined with the other measures that constitute the evaluation framework (that is, with building-level data, student growth data, and elective data). We conducted three analyses to assess the extent of variation in rubric scores. First, we tabulated the percentages of Phase 2 teachers who were rated in each of the four performance categories by rubric component to see how often principals used each category. Second, given those distributions of ratings, we next examined whether the observed variation in rubric scores is comparable to the amount of variation in Phase 2 teachers' estimated contributions to their students' achievement gains as measured by value added.[14] Third, because variation in teachers' scores is meaningful only if based on consistent applications of the rubric by principals, we investigated the extent to which principals' ratings in Phase 2 varied within versus between schools, and explored possible reasons why ratings might differ systematically across schools.

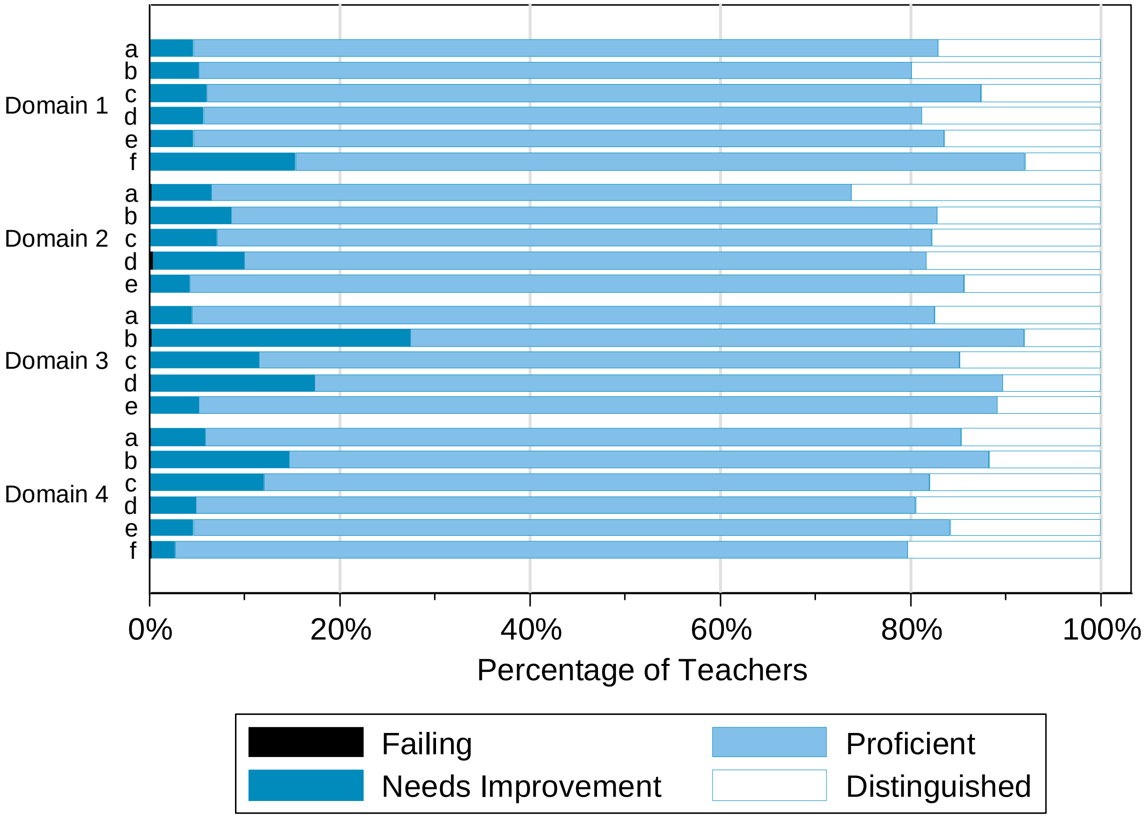
A. Most Teachers Received Positive Ratings on Phase 2 Components and Summary Measures, Though There Were Differences by District

Although exact percentages varied by component, typically the practices of at least 90 percent of teachers were rated as either *proficient* or *distinguished* on components of their professional practices (Figure III.1; see Appendix Table A.1 for tabular format). The most common rating was *proficient,* which included between 65 and 84 percent of teachers. About 8 percent of teachers typically received a *needs improvement* rating, and less than 1 percent typically received a *failing* rating.

The extent of positive ratings on Phase 2 components compares with findings for Phase 1 (Lipscomb et al. 2012) but it contrasts with the MET study (Kane and Staiger 2012). For example, the most difficult Phase 2 component (as measured by based on the largest proportions of teachers with *needs improvement* or *failing* ratings) was component 3b, which measures questioning and discussion techniques. Whereas 27 percent of Phase 2 teachers received one of the two lowest ratings on component 3b (including 0.2 percent rated as *failing*), over half of teachers in the MET study were given the second-lowest rating on the comparable component, and about 7 percent were given the lowest. The Phase 2 component with the largest proportion of teachers rated *distinguished* was component 2a—creating a learning environment of respect and rapport. Whereas 7 percent of Phase 2 teachers received a rating below *proficient* on this component, about 25 percent of teachers received a low rating on a similar component in the MET study. There are several possible explanations for these differences. For instance, the MET study referred to the lowest two categories as *basic* and *unsatisfactory,* which may have more positive connotations than *needs improvement* and *failing* (though Pittsburgh also uses *basic* for the second-lowest performance category). Other explanations could include differences in training provided to observers and differences between the studies in the effectiveness of participating teachers.

[14] See Appendix B for technical details on the VAMs.

Figure III.1. Summary of Phase 2 Rubric Component Scores—All Districts

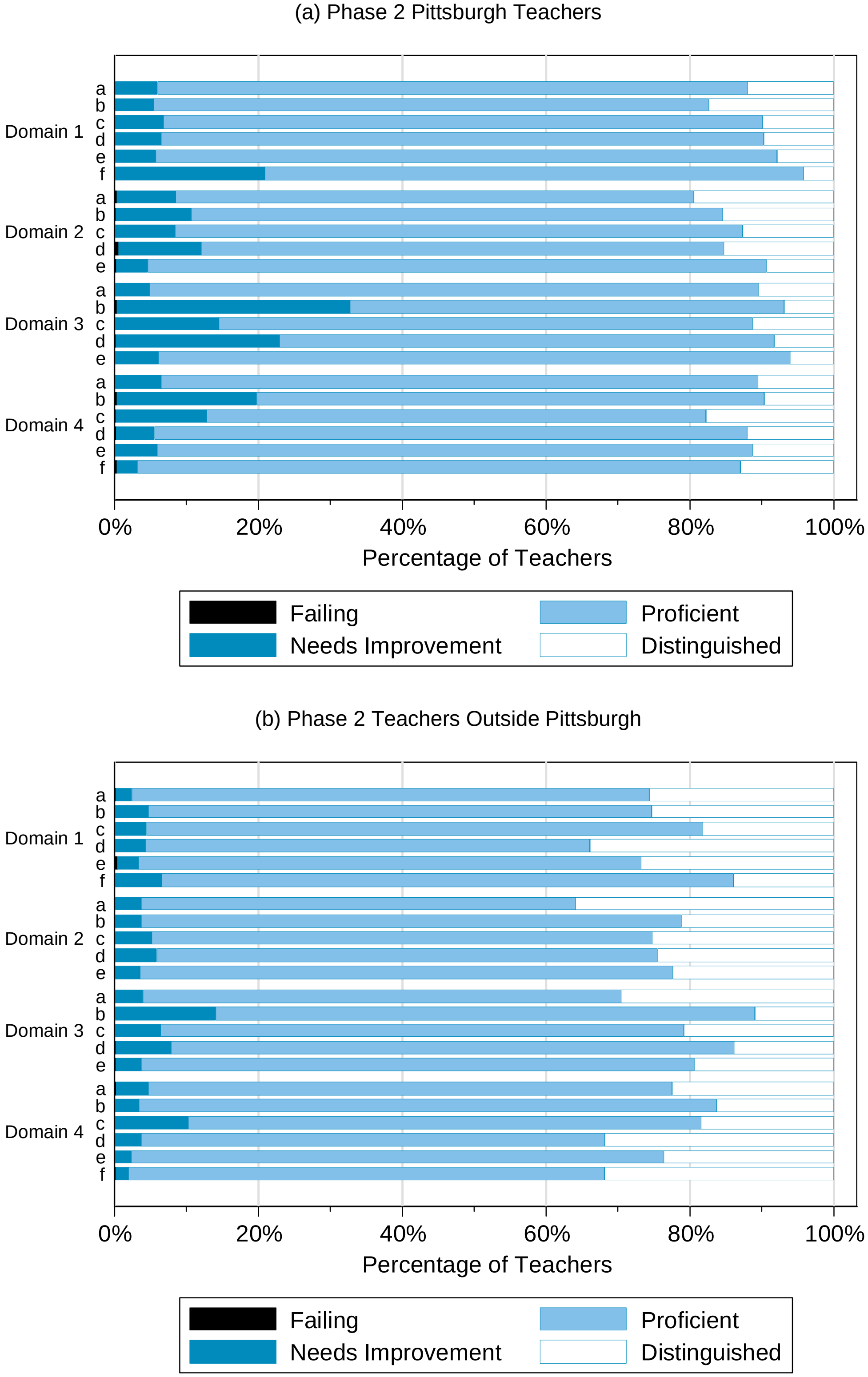


Source: Mathematica calculations based on Phase 2 classroom observation data from 2011–2012.

Note: We report sample sizes by component in Appendix Table A.1.

Teachers in Pittsburgh tended to receive lower scores on the rubric than teachers outside of Pittsburgh. Figures III.2a and III.2b show the distribution of scores by component for teachers in each sample. In Pittsburgh, the proportion receiving a *distinguished* rating was lower for every component. On the average component, 11 percent of Pittsburgh teachers received a *distinguished* rating versus 23 percent for Phase 2 teachers outside Pittsburgh. The proportion of teachers receiving a *needs improvement* rating was higher in Pittsburgh for all components but 4a (reflecting on teaching and student learning) and 4f (showing professionalism). On average, the professional practices of 10 percent of Pittsburgh teachers were rated as *needs improvement,* compared to 5 percent outside Pittsburgh. Appendix Tables A.2 and A.3 show the information depicted in Figure III.2 in table form.

Figure III.2. Summary of Phase 2 Rubric Component Scores—Pittsburgh and Non-Pittsburgh Samples

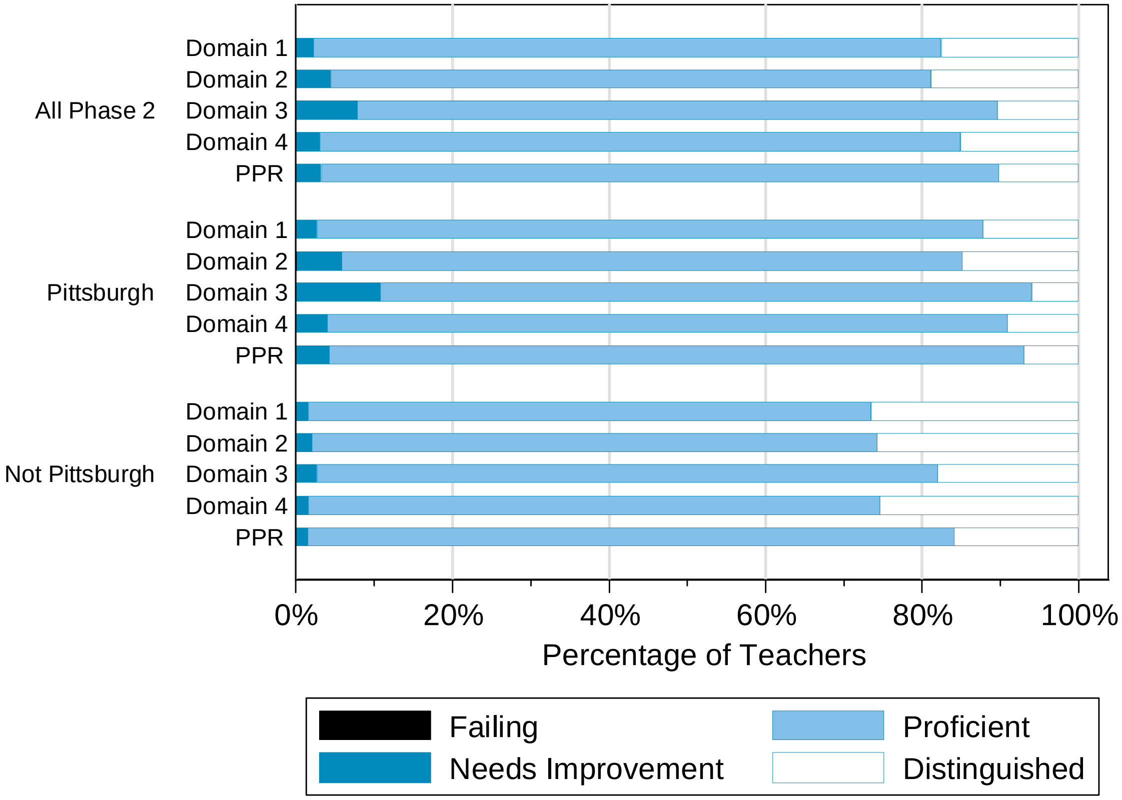


Source: Mathematica calculations based on Phase 2 classroom observation data from 2011–2012.

Note: We report sample sizes by component for Pittsburgh in Appendix Table A.2 and for teachers in districts other than Pittsburgh in Appendix Table A.3.

The lower scores in Pittsburgh were also reflected in the domain-level scores and the PPR. In Figure III.3, the scores on these measures are grouped into the four performance categories, with scores of less than 0.5 scored as *failing,* at least 0.5 but less than 1.5 as *needs improvement,* at least 1.5 but less than 2.5 as *proficient,* and at least 2.5 as *distinguished.* As with the component scores, Pittsburgh teachers more frequently received *needs improvement* ratings on the domain averages and on the PPR than did teachers outside Pittsburgh. No teachers received a *failing* score on any of the domain-level scores or the PPR.

Figure III.3. Summary of Scores on Rubric Domains and Professional Practice Ratings by District



Source: Mathematica calculations based on Phase 2 classroom observation data from 2011–2012.

Note: Scores of less than 0.5 are considered *failing,* scores of at least 0.5 but less than 1.5 are considered *needs improvement,* scores of at least 1.5 but less than 2.5 are considered *proficient,* and scores of at least 2.5 are considered *distinguished.*

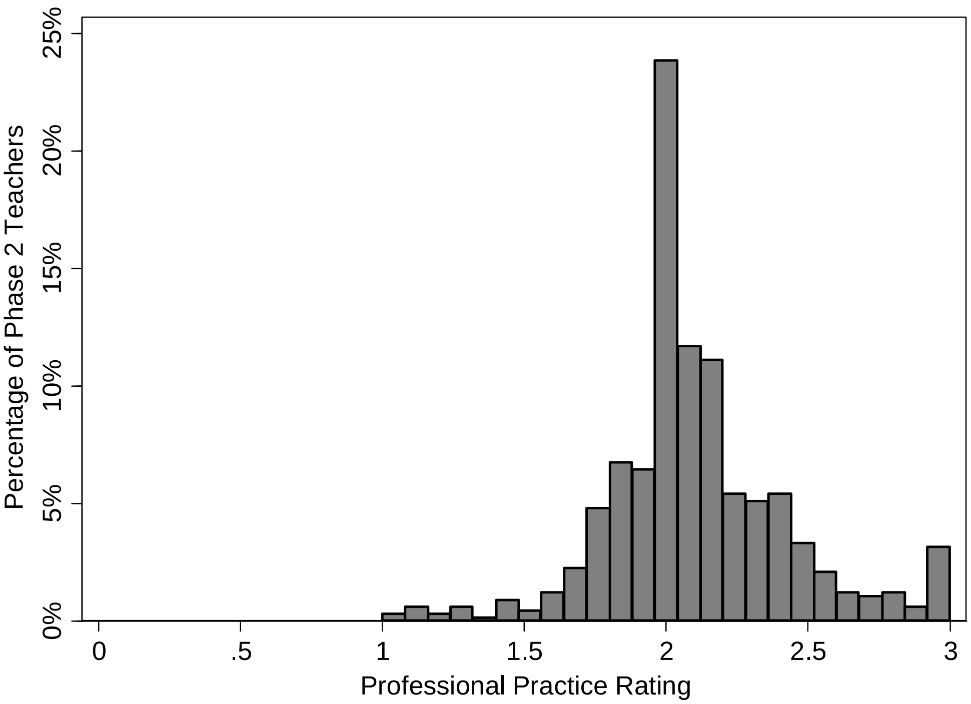
PPR = Professional Practice Rating.

B. Summary-Level Rubric Ratings Have More Potential to Differentiate Teacher Effectiveness than Individual Components, but Value Added May Differentiate the Most

1. Distribution of Teachers' Professional Practice Rating Scores

Although most summary-level rubric ratings (that is, domain-level ratings and the PPR) are concentrated within the *proficient* category (Figure III.3), these ratings have more potential to distinguish between teachers' performance than any individual component. Whereas each individual rubric component distinguishes teachers into one of four categories, summary-level scores produce finer distinctions. Figure III.4 demonstrates the extent of variation in scores on the PPR. The height of each bar gives the fraction of teachers with scores on the PPR as indicated on the horizontal axis. The tallest bars fall between scores of 1.5 and 2.5, the range of scores for teachers who are rated *proficient*. Not all scores distinguish teachers—almost 25 percent of Phase 2 teachers received a score of approximately 2. Similar rubric scores are grouped into bins of teachers, so even within the bars shown in the figure, teachers have distinct PPR scores. For example, though the tallest bar includes some teachers scoring just above or below 2, 17 percent of teachers received a score of exactly 2.

Figure III.4. Distribution of Professional Practice Ratings—Phase 2 Teachers with Value Added



Source: Mathematica calculations based on Phase 2 classroom observation data from 2011–2012 and value-added estimates from 2009–2010 to 2011–2012.

2. Comparison of Variation in Value-Added Estimates and Professional Practice Rating Scores

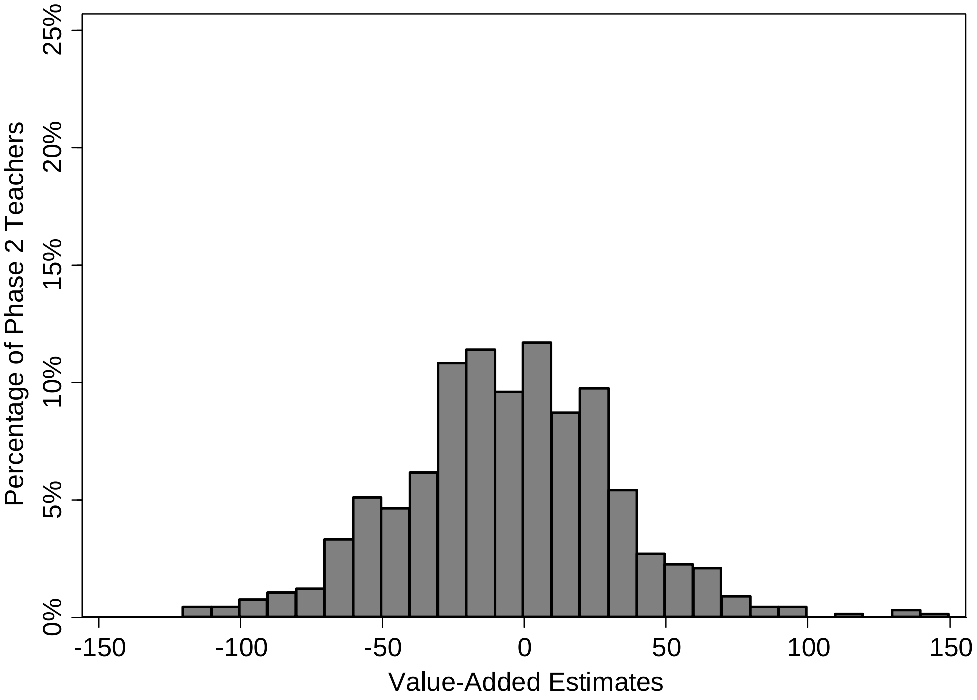
Summary-level rubric ratings have the potential to distinguish between teachers in terms of their performance to a similar degree as other teacher effectiveness measures, such as value-added estimates—measures of teachers' contributions to their students' achievement gains based on standardized assessment scores. However, as we describe below, because the reliability of value-added estimates is likely higher than the reliability of professional practice rating scores, value added has the most potential to differentiate teacher effectiveness.

Value-added estimates fall on a continuous scale and are expressed in terms of the number of test points larger or smaller than the magnitude of the contribution of the average teacher in the sample. There are few, if any, ties, and so value added has a lot of potential to distinguish teacher effectiveness. However, value-added estimates also are measured with some amount of imprecision—estimation error in teachers' scores resulting from the typically small numbers of students taught, and the variety of factors that affect student achievement besides teacher effectiveness—so small differences may not be meaningful. This amount of imprecision can be measured using a confidence interval.[15]

Like value added, summary-level rubric scores also fall on a continuous scale, though the scale is bounded between zero and three. Unlike value added, it is not typical to report rubric scores along with a confidence interval, though, like value added, they are certain to be measured with some degree of imprecision. This imprecision arises because rubric scores are assigned based on a limited amount of time in the classroom and a limited number of components, and are conducted by a single principal. Consequently, a teacher's actual effectiveness, if measured without any imprecision, may be likely to fall within some fixed range around the score the teacher was actually assigned.

Visual inspection of the distributions of PPR scores and value-added scores (Figures III.4 and III.5, respectively) suggests that value added might provide more distinctions in the center of the distribution, though such visualizations do not account for the imprecision in value-added estimates or rubric scores.[16] As in Figure III.4, Figure III.5 groups similar value-added estimates into bins of teachers. Unlike in the distribution of PPR scores, the largest proportion of teachers in any single bar for value added is about 12 percent. The distribution of value-added estimates is more even in the center of the distribution, potentially allowing for more distinctions between teachers depending on the amount of imprecision in the estimates.

Figure III.5. Distribution of Value-Added Estimates—Phase 2 Teachers with Value Added



Source: Mathematica calculations based on Phase 2 classroom observation data from 2011–2012 and value-added estimates from 2009–2010 to 2011–2012.

[15] Appendix Table C.1 contains descriptive statistics on the value-added estimates, including a measure of precision.

[16] The sample for Figure III.4 and Figure III.5 is the same (that is, Phase 2 teachers with value-added estimates).

A comparison of the reliability of rubric scores and value-added estimates suggests that value added has more potential to differentiate teacher effectiveness. Whereas confidence intervals from the VAM provide a direct measure of imprecision in value-added estimates, confidence intervals are not available for the rubric. In Section II we found that the PPR had an internal consistency of 0.84. This value likely overstates the reliability of the PPR, because it does not account for the contributions of specific raters and observed lessons to imprecision. The MET study (Kane and Staiger 2012; Ho and Kane 2013) has found reliabilities of under 0.6 for similar rubrics that do account for these other factors. In contrast, the confidence intervals for value-added estimates imply that the reliability of value added is 0.88. Consequently, value-added may make finer distinctions among teachers than the rubric.

C. Teacher Performance on the Rubric Varies Mostly Within Schools, but It Also Varies Between Schools, Which Suggests that Continued Evaluator Training Will Be Important

For variation across teachers in their rubric scores to be meaningful in an evaluation system, different evaluators must apply the rubric consistently in conducting their observations. If instead the variation in rubric scores is a product of equally effective teachers receiving different ratings because the evaluators applied different standards, then there would be a clear need for further evaluator training, as rubric scores would not be comparable across teachers. We explored this issue in two ways. First, we examined whether Pittsburgh teachers, shown in Figure III.2 to have lower rubric ratings than other Phase 2 teachers, also tended to have lower value-added estimates. If principals applied the rubric consistently, then the lower rubric scores for Phase 2 teachers in Pittsburgh should also be reflected, at least in part, in lower value-added estimates.

Even if there is some information to suggest that a portion of the differences in rubric scores is due to differences in value-added estimates, another portion still could be due to principals applying the rubric differently across schools within these groups. Therefore we conduct a second analysis to examine the proportion of variation in rubric scores that is explained by between-school versus within-school factors, and then to compare it to the similar proportions for value-added scores. Between-school variation in rubric scores (and value added) could be due to differences in effectiveness or to differences in the sample of Phase 2 teachers by school. Some schools participating in Phase 2 may be better at attracting, training, and retaining effective teachers for a number of reasons, in which case we would expect there to be between-school variation in teachers' rubric ratings. It is also possible that teachers were selected to participate in Phase 2 in different ways across schools based on their effectiveness, which could lead to between-school variation in rubric ratings as well. Because the characteristics of teachers varied across districts in the Phase 2 sample—for example, a lower proportion of teachers had a master's degree in Pittsburgh than in the non-Pittsburgh Phase 2 sample—the possibility of differential selection of participating teachers across schools cannot be discounted. Another possibility is that principals apply the rubric in different ways, and this leads to differences in average rubric scores between schools.

A typical finding in the research literature on value added is that teacher effectiveness varies more within schools than between them (see, for example, Lipscomb et al. 2012). In other words, there are fewer differences in the average effectiveness of schools than there are in the effectiveness of teachers in each of the schools. We expected to see a similar pattern for rubric scores, where there is less variation between schools than within schools because our priors are both that between-school differences are smaller and that principals are applying the rubric consistently. We also expected to see that the proportion of variation in rubric scores that is due to between-school factors corresponds to the proportion of variation in value-added scores that is due to between-school factors. If there is more between-school variation in rubric scores than in value-added scores, it could be an indication that principals are applying the rubric in different ways.

Although we cannot confidently distinguish between explanations for differences in ratings across districts and schools using the Phase 2 data, we found that (1) Phase 2 teachers in Pittsburgh did have lower value-added estimates on average, and this difference is comparable to the difference in rubric scores; and (2) there may be a higher-than-expected amount of between-school variation in rubric scores, consistent with an interpretation that principals may have applied the rubric differently. We believe the prudent recommendation based on the available information is for continued evaluator training, especially while the system is still new.

1. Differences in Teacher Effectiveness Ratings Across Districts

In Table III.1, we report average domain-level rubric scores and PPR scores separately for Phase 2 teachers in Pittsburgh and in other districts. In the last row, we show average value-added estimates for the teachers in each group that have estimates. We note that these comparisons are based on the Phase 2 sample, rather than all teachers in the state. For example, the smaller average scores in Pittsburgh may not characterize the effectiveness of Pittsburgh teachers overall relative to other urban districts or to the state average if the rubric were implemented at scale.

Table III.1. Average Scores by Rubric Domains and Professional Practice Rating

	All Phase 2 Teachers		Phase 2 Teachers with VAM	
	Pittsburgh	Not Pittsburgh	Pittsburgh	Not Pittsburgh
Domain 1: Planning and Preparation	2.0	2.2*	2.1	2.2*
Domain 2: Classroom Environment	2.1	2.2*	2.1	2.2*
Domain 3: Instruction	1.9	2.1*	2.0	2.1*
Domain 4: Professional Responsibilities	2.0	2.2*	2.0	2.2*
Professional Practice Rating	2.0	2.2*	2.1	2.2*
Value-Added Estimates (In generalized PSSA scaled score units)	n/a	n/a	−13.9	−2.4*
Number of Phase 2 Teachers	1,667	948	395	271

Source: Mathematica calculations based on Phase 2 classroom observation data from 2011–2012 and value-added estimates from 2009–2010 to 2011–2012.

Note: The domain-level scores are calculated for teachers with observation data on at least one component in the domain. The Professional Practice Rating is calculated for all Phase 2 teachers, reweighting the domains in the same proportions in the event that a teacher does not have any components rated in a domain. The number of teachers with scores for a particular domain-level average is smaller than the number of Phase 2 teachers.

* Mean difference between Pittsburgh and other teachers is statistically significant at the 5 percent level.

The data suggest a relationship between lower value-added estimates in Pittsburgh and the tendency of Pittsburgh teachers to receive lower rubric ratings. Pittsburgh teachers in Phase 2 with value-added estimates scored, on average, 0.1 rubric unit lower than Phase 2 teachers in other districts and 11.5 PSSA points lower in terms of value added.[17] To put these differences into

[17] As described in Appendix B, we report value added in terms of "generalized" PSSA scaled score points—points on a typical PSSA assessment in any grade or subject. Appendix B also describes our method for combining value-added estimates across grades and subjects to obtain the teacher-level estimates that we ultimately use in the analysis.

perspective, 11.5 points means that the average Phase 2 teacher outside Pittsburgh contributed an additional 11.5 points to the achievement of his or her students on a typical PSSA assessment. This contribution is equivalent to moving a teacher at the median of the effectiveness distribution to the 60th percentile. In contrast, an improvement of 0.1 on the PPR is equivalent to moving a teacher at the median of the professional practice distribution to the 63rd percentile. The similar magnitudes of these two differences suggest that differences in teacher effectiveness account for much of the difference in rubric scores between Pittsburgh and other teachers during Phase 2.

2. The Proportion of Variation in Rubric Scores Explained by Between-School Factors

Although rubric scores may differ across schools and districts because the value-added scores of teachers in those schools and districts vary, it could also be that principals vary in how they apply the rubric. We examine this possibility in Table III.2, by showing the proportion of variation in Phase 2 rubric scores that is between schools. The first set of columns shows separate results for all Phase 2 teachers, Pittsburgh teachers, and other Phase 2 teachers. The second set shows results for each of these groups among teachers with value-added estimates. The rows pertain to each domain and to the PPR. The last row shows comparable information for value added among teachers with those scores.

As expected, most, but not all, of the variation in rubric ratings is within schools rather than between them. Overall, 38 percent of the variation in PPR scores is between schools (column 1), or 41 percent if we restrict to Phase 2 teachers with value-added estimates (column 4). The percentages of between-school variation are lower in Pittsburgh than elsewhere. This may be expected because the non-Pittsburgh sample includes a smaller number of teachers per school from more schools across many districts in the state. A smaller number of teachers per school will tend to produce more between-school variance even if there is no difference in how teachers are distributed across schools based on their effectiveness.

Regardless of the sample, the between-school proportions of variation in rubric scores were larger than the same proportions for value added, though these differences might not be statistically significant. The differences for PPR scores were 5 percentage points in Pittsburgh and 15 outside Pittsburgh.[18] If principals were applying the rubric in exactly the same ways across schools, then we would expect to see, at most, small differences in these percentages.[19] The fact that we see larger proportions of between-school variation in rubric ratings suggests that principals may not be applying the rubric in exactly the same ways. That this difference between rubric ratings and value added is largest outside Pittsburgh provides additional support for this possibility, because principals in Pittsburgh have more experience with the rubric, and thus have greater familiarity with how it is intended to be applied. However, 5—or even 15—percentage points might not be a statistically significant difference, and there may be other explanations for this difference.

[18] The proportion of between-school variation in value-added estimates for Phase 2 is comparable to findings for Pennsylvania as a whole during Phase 1 (Lipscomb et al. 2012).

[19] Some differences in the amount of between-school variation could be caused by differences in the amount of imprecision between the rubric and value added.

Table III.2. Proportion of Variance in Rubric Domains and Value-Added Between Schools

	Proportion of Variance Between Schools					
	All Phase 2 Teachers			Phase 2 Teachers with VAM		
	All	Pittsburgh	Not Pittsburgh	All	Pittsburgh	Not Pittsburgh
Domain 1: Planning and Preparation	0.37	0.18	0.53	0.41	0.32	0.57
Domain 2: Classroom Environment	0.31	0.17	0.47	0.39	0.25	0.62
Domain 3: Instruction	0.37	0.22	0.52	0.41	0.30	0.61
Domain 4: Professional Responsibilities	0.35	0.17	0.53	0.41	0.29	0.55
Professional Practice Rating	0.38	0.22	0.57	0.41	0.30	0.64
Value-Added Estimates	n/a	n/a	n/a	0.35	0.25	0.49
Number of Phase 2 Teachers	2,571	1,667	904	606	395	211
Number of Schools	268	64	204	124	53	71

Source: Mathematica calculations based on Phase 2 classroom observation data from 2011–2012 and value-added estimates using data from 2009–2010 to 2011–2012.

Notes: We exclude schools with only one teacher from this analysis.

The domain-level scores are calculated for teachers with observation data on at least one component in the domain. The Professional Practice Rating is calculated for all Phase 2 teachers, reweighting the domains in the same proportions in the event that a teacher does not have any components rated in a domain. The number of teachers with scores for a particular domain-level average is smaller than the number of Phase 2 teachers.

Value-added estimates are post shrinkage.

Our findings suggest that differences in teacher effectiveness account for much of the difference in rubric scores between Pittsburgh and other Phase 2 teachers. We also found evidence that suggests there is more between-school variation in rubric ratings than in value added, especially in the non-Pittsburgh sample. While this evidence does not prove that principals vary in how they apply the rubric, it is one possible explanation for the differences. In particular, one conclusion supported by these data is that the greater exposure that Pittsburgh principals had to the rubric by implementing it over two years led to smaller differences across Pittsburgh schools than across schools in other districts in how the rubric is being applied. Based on this possible explanation, we recommend that PDE continue its efforts to train evaluators and follow up on that training periodically, especially since the rubric is still new to principals.

IV. RELATIONSHIPS BETWEEN RUBRIC SCORES AND VALUE ADDED

PDE views the measurement of effective teaching practices using the rubric as an intermediate output in a logic model for the production of student learning (Figure I.1). If the rubric indeed measures teacher practices that are related to student outcomes, then teachers with higher rubric scores should tend to score higher on indicators of teaching effectiveness that are directly based on student outcomes. We compared teachers' rubric scores to their contributions to student achievement estimated using a VAM. Strong relationships between rubric scores and value added can validate the logic model that underpins the rubric. Therefore, the strength of the relationship between the rubric scores and value-added scores is a measure of the validity of the rubric.

An advantage of the rubric is that it measures a wide range of teaching practices. Individual practices may vary in terms of the degree to which they relate specifically to teachers' contributions to their students' achievement gains as measured by a VAM. Although we expect that some components will have stronger associations than others with value added, a higher score on any component that adheres to PDE's logic model should translate to some degree into above-expected contributions to student achievement growth. This should be true especially for the PPR, since it will be used directly in calculating teachers' final evaluation scores. The rubric components that have the largest positive associations with value added could be promising practices for PDE to target for professional development. PDE also might consider assigning more weight to rubric components or domains that are more strongly associated with value added in the final evaluation system.

After examining the validity of rubric components combining all grades and subjects covered by our VAM, we estimate relationships separately for specific grade spans and subjects. These findings could indicate whether variations in the observation rubric by grade and subject might be recommended. A limitation of the VAM is that the analysis can include only 4th- through 8th-grade teachers in tested subjects. Findings from our analyses could be suggestive of the validity of rubric components in other grades and subjects, but we cannot extend our findings directly. For example, if the strength of associations between value added and rubric components are similar across all grades and subjects included in the analysis, this would support the contention that the same weights could be used to construct the PPR for teachers in other grades and subjects.

Finally, we conduct two types of robustness checks. First, we ask whether the key findings are likely to be due to the characteristics of the particular sample of teachers that participated in the Phase 2 pilot. This robustness check attempts to better understand how similar we would expect the results to be had all teachers in the state participated. Second, we explore whether the results are affected by changing the student cohorts used in the VAM. Our primary results use a VAM that combines the 2011–2012 cohort of students—the year of the Phase 2 rubric evaluations—with the prior two cohorts, because this approach best reflects the VAM that PDE is likely to use in the final evaluation system. Including student achievement data from 2011–2012, however, could overstate the degree to which rubric scores correspond to value-added estimates because some of the same students will contribute to both measures. However, excluding student achievement from this year could understate these correlations, because there could be real differences in teacher effectiveness across years. Therefore, we also estimate a three-cohort VAM that includes 2008–2009 to 2010–2011 student cohorts, but not the 2011–2012 cohort. Similarly, we estimate one-cohort VAMs that include only 2011–2012 students and only 2010–2011 students, respectively. Although the VAMs that do not include 2011–2012 may provide for correlations that are lower in magnitude, we do not interpret them as providing lower-bound validity estimates when Pittsburgh teachers are included in the sample. Pittsburgh principals were provided value added information about their schools and teachers following the 2010-2011 year, which may have factored into their subjective ratings of teachers in 2011–2012, thereby increasing correlations with any of the four sets of VAM estimates as well (Rockoff et al. 2010).

A. Most Elements of the Rubric Are Related to Value Added

To assess validity, we correlated value-added estimates with each rubric component, domain average, and the professional practice rating. The possible range of values for correlations is from −1.0 to 1.0. A positive correlation indicates that a higher rubric score is associated with a higher value-added estimate. The closer the correlation is to 1.0 in absolute value, the larger the association. A correlation of 0 indicates no association. We estimate each correlation using a regression model and then scale the coefficient estimate to obtain the correlation. The correlations we report can be compared to similar correlations in other studies, including those in the MET studies.

As is frequently done by other researchers—in the MET studies, for example—we adjust the correlations to reflect the correlation with *underlying value added*—the measure of value added we would obtain if we could eliminate estimation error.[20] Value-added estimates are measured with some imprecision, which tends to lower correlations with the rubric. By making this adjustment, we can correlate rubric scores more directly with the portion of value-added scores that is signal rather than noise (that is, estimation error).

In Figure IV.1, we depict the relationship between teachers' PPR scores and their value-added estimates. Each dot in the figure is an individual Phase 2 teacher with a value-added estimate. The upward sloping line in the figure indicates that the correlation is positive—so a higher value-added estimate is associated with a higher PPR. A plot with more teachers close to this line would produce a correlation closer to 1.0, whereas a plot with teachers scattered across the chart would produce a correlation closer to zero. As reported in the first column of Table IV.1, this correlation is 0.24.

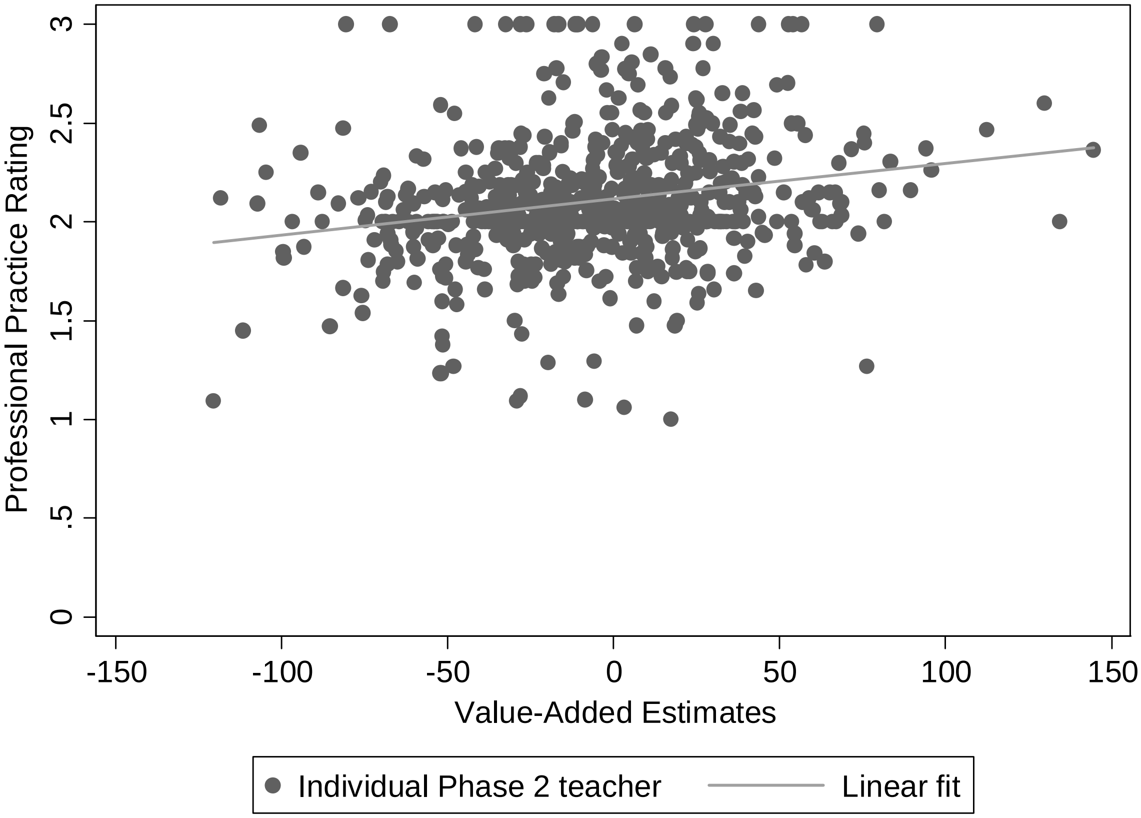
With a small, positive correlation like 0.24, we expect that some teachers will have low value-added estimates and high PPR scores, while others will have high estimates and low scores. For example, 17 percent of Phase 2 teachers with value-added scores received a PPR score of exactly 2. These teachers' value-added estimates ranged from −127 to +136. Eight Phase 2 teachers had value-added estimates within 1 unit of a zero value. These latter teachers' PPR scores ranged from 1.1 to 2.9.

We found correlations between rubric scores and value added that were in a range consistent with prior research using the Framework for Teaching rubric (for example, the MET study by Kane and Staiger 2012). The correlations with each rubric domain and the PPR based on all Phase 2 teachers ranged from 0.17 to 0.28 and were all statistically significant (Column 1 of Table IV.1); that is, we can say with confidence that the correlations are positive. Domain 3—instruction—had the largest correlation with value added, and domain 4—professional responsibilities—had the smallest correlation. Correlations with individual components ranged from 0.04 (organizing physical space) to 0.25 (communicating with students) and were statistically significant, except for component 2e (organizing physical space) and 4f (showing professionalism). Among components, larger correlations tended to be found in the instruction domain, and lower correlations tended to be

[20] We adjust the correlations by scaling them by the inverse of the square root of the estimated reliability of the value-added estimates. We calculate this reliability using the estimated standard errors on the value-added estimates. Jacob and Lefgren (2008) describe this method.

found in the professional responsibilities domain. Among Phase 2 teachers in Pittsburgh and elsewhere, the summary level correlations were for the most part similar. At the component level, the correlations were larger for Pittsburgh teachers on some components and smaller on other components. Among the three components that were to be rated for all teachers (that is, 1e, 3c, and 3d), the correlations for the Phase 2 sample were 0.18, 0.22, and 0.17 respectively, and each was statistically significant.

Figure IV.1. Relationship Between Professional Practice Ratings and Value Added



Source: Mathematica calculations based on Phase 2 classroom observation data in the 2011–2012 school year and value-added estimates using data from 2009–2010 through 2011–2012.

Our correlation results by domain are consistent with PDE's plan to place more weight on the instruction domain than the professional responsibilities domain in calculating the PPR. However, we find larger correlations in the planning and preparation domain than in the classroom environment domain, though the former is currently given less weight in the PPR.

Table IV.1. Validity of Rubric Components, Domains, and Professional Practice Rating

	All Phase 2		Pittsburgh		Not Pittsburgh	
	Correlation	Teachers	Correlation	Teachers	Correlation	Teachers
Domain 1: Planning and Preparation	0.23*	637	0.20*	369	0.21*	268
Domain 2: Classroom Environment	0.19*	637	0.18*	372	0.16*	265
Domain 3: Instruction	0.28*	660	0.27*	389	0.24*	271
Domain 4: Professional Responsibilities	0.17*	633	0.16*	368	0.11	265
Professional Practice Rating	0.24*	666	0.22*	395	0.22*	271
1a: Demonstrating knowledge of content and pedagogy	0.12*	451	0.06	262	0.17*	189
1b: Demonstrating knowledge of students	0.19*	585	0.18*	368	0.18*	217
1c: Setting instructional outcomes	0.14*	575	0.10	368	0.19*	207
1d: Demonstrating knowledge of resources	0.17*	444	0.15*	262	0.15	182
1e: Planning coherent instruction	0.18*	525	0.19*	263	0.13*	262
1f: Designing ongoing formative assessments	0.16*	451	0.16*	260	0.10	191
2a: Creating a learning environment of respect and rapport	0.14*	496	0.16*	272	0.09	224
2b: Establishing a culture for learning	0.20*	571	0.18*	371	0.21*	200
2c: Managing classroom procedures	0.18*	482	0.24*	269	0.10	213
2d: Managing student behavior	0.16*	587	0.17*	368	0.12	219
2e: Organizing physical space	0.04	453	0.01	260	0.04	193
3a: Communicating with students	0.25*	440	0.28*	262	0.20*	178
3b: Using questioning and discussion techniques	0.24*	551	0.24*	372	0.21*	179
3c: Engaging students in learning	0.22*	639	0.21*	372	0.19*	267
3d: Using assessment to inform instruction	0.17*	643	0.19*	376	0.08	267
3e: Demonstrating flexibility and responsiveness	0.18*	435	0.17*	259	0.17*	176
4a: Reflecting on teaching and student learning	0.13*	605	0.14*	367	0.08	238
4b: System for managing students' data	0.09*	574	0.06	368	0.09	206
4c: Communicating with families	0.12*	555	0.15*	367	0.00	188
4d: Participating in a professional community	0.13*	450	0.07	261	0.15*	189
4e: Growing and developing professionally	0.12*	438	0.09	257	0.11	181
4f: Showing professionalism	0.10	401	0.03	222	0.12	179

Source: Mathematica calculations based on Phase 2 classroom observation data in the 2011–2012 school year and value-added estimates using data from 2009–2010 through 2011–2012.

Note: Correlations are adjusted for estimation error in value-added estimates (Jacob and Lefgren 2008). An unadjusted correlation is 7 percent smaller on average.

* Statistically significant at the 5 percent level.

B. Across Grade Spans and Subjects, Relationships Between Rubric Scores and Value Added Are Typically Positive, Though Magnitudes Can Vary

We next conducted separate analyses for math, ELA, and science departmentalized teachers in grades 6 through 8, and for homeroom teachers—teachers responsible for teaching more than a single subject—in grades 4 through 6. Comparing the correlations across these groups could indicate whether variations in the observation rubric by grade and subject might be recommended. For example, if certain rubric components are associated with higher value-added scores in some, but not all, grades and subjects, then PDE may consider recommending that principals use these components especially for teacher evaluations where the associations are strongest. Alternatively, if a component is associated with higher value-added scores in all grades and subjects, PDE may have more confidence in using it to assess teachers in non-tested grades and subjects.

Correlations by grade span and subject are reported in Table IV.2 by domain and for the PPR. We find that the correlations for teachers in grades 4 through 6 who teach multiple subjects tended to be smaller than the correlations for departmentalized teachers in grades 6 through 8. The main exception was in domain 4 (professional responsibilities), which had smaller correlations for departmentalized math and ELA teachers. Relative to Table IV.1, fewer of the correlations are statistically different from zero once the sample is partitioned into multiple groups. The lack of statistical significance in some table cells is likely due to the lower sample sizes; indeed, the precision of several of the correlations like the PPR for homeroom teachers is very near the limit of what can be considered statistically significant using a 95 percent confidence interval.

Table IV.2. Validity of Rubric Domains and Professional Practice Rating by Subject and Grade Span—All Districts

	Grades 4 Through 6 Homeroom Teachers		Grades 6 Through 8 Departmentalized Teachers of:					
			Math		ELA		Science	
	Corr.	Teachers	Corr.	Teachers	Corr.	Teachers	Corr.	Teachers
Domain 1: Planning and Preparation	0.08	130	0.28*	115	0.24*	162	0.21	48
Domain 2: Classroom Environment	0.15	132	0.13	116	0.21*	161	0.32*	45
Domain 3: Instruction	0.17*	134	0.29*	120	0.20*	171	0.56*	48
Domain 4: Professional Responsibilities	0.18*	132	−0.03	115	0.03	158	0.36*	47
Professional Practice Rating	0.17	134	0.22*	121	0.16	172	0.46*	48

Source: Mathematica calculations based on Phase 2 classroom observation data in the 2011–2012 school year and value-added estimates using data from 2009–2010 through 2011–2012.

Notes: Correlations are adjusted for estimation error in value-added estimates (Jacob and Lefgren 2008).

We include the 15 percent of grade 6 teachers who teach both math and ELA as homeroom teachers in the columns for elementary grade teachers. Total sample sizes in each table row are less than in Table IV.1 because teachers in grades 7 and 8 are excluded if they teach multiple subjects.

* Statistically significant at the 5 percent level.

Correlations for departmentalized science teachers in grades 6 through 8 were larger than in other subjects or grades for the PPR, and for all domains except for planning and preparation. For example, the correlation with the PPR for science teachers was 0.46, whereas the largest correlation in the other grades and subjects was 0.29 in math. This result could indicate that the rubric is a stronger indicator of a teacher's contributions to student achievement growth in science than in other subjects or grades.[21] If so, PDE might consider placing additional weight on the rubric for departmentalized science teachers. However, because the correlations for science are based on data from fewer than 50 teachers, we recommend that this finding be confirmed based on the larger sample that is expected for Phase 3. Regardless, our finding of correlations that are consistently numerically positive (if not always statistically positive, as a result of smaller samples) across subject and grade levels suggests that the rubric might also be associated with student achievement growth in non-tested grades and subjects.

C. Accounting for Principals' Selection of Teachers Did Not Substantively Change Correlation Estimates

Correlations between rubric scores and value added based on the Phase 2 sample may not reflect the correlations we would obtain had all teachers in the state participated, or had the teachers been selected randomly. The Phase 2 sample is not representative of teachers in Pennsylvania as a whole (Table I.1). Although most teachers in Pittsburgh received observation scores, principals outside Pittsburgh selected which of their teachers to observe for the Phase 2 rubric. Principals in all districts were advised not to select teachers who had been rated as unsatisfactory in previous years. Consequently, the correlations we estimate using the Phase 2 sample apply only to the teachers who actually participated and not to teachers in the state generally. This is less of a concern for Phase 2 teachers in Pittsburgh, because most teachers in this district participated.[22]

In addition to producing a nonrepresentative sample, nonrandom selection of teachers can lead to possible bias in the correlation estimates if teachers are selected based on characteristics related to student achievement gains. For example, if principals selected only highly effective teachers (according to value added) for Phase 2, correlations between rubric scores and value added might be lower than if the pilot had included a random sample of teachers. This is because among the teachers with low rubric scores, only those with high value added are selected in this example. In the data, it would appear that teachers with low rubric scores have higher value-added scores than is the case for the typical teacher in the state, which leads to a weaker correlation. The correlations could also be biased if principals differ in how they select teachers. One principal might select the most-effective teachers in the school for participation and tend to give low scores on the rubric, while another might select only the least-effective teachers and tend to give high scores on the rubric. In this case, comparing the teachers across these schools could result in a spurious *negative* relationship between the rubric and value-added scores. Although in this example correlations that do not adjust for selection may be too low, another scenario could lead to the opposite bias. The correlations in Table IV.1 could be too low or too high because of selection bias.

[21] We also calculated correlations in math and reading based on teachers who have students only in grade 8, because value added for departmentalized science is based solely on grade 8. Correlations in grade 8 ELA were similar to correlations for all middle school ELA teachers, but correlations for grade 8 math were similar to those for science. Because grade 8 sample sizes were small, we recommend that these findings be confirmed based on the larger sample that is expected for Phase 3.

[22] Though participation in Pittsburgh was high, many teachers were not rated on all rubric components. Consequently, selection of components could lead to bias in the component-level correlations with value added. Consistent with the results for domain-level scores and the PPR presented in this section (Table IV.5), accounting for selection leads to small to moderate decreases in correlations between individual rubric components and value added in Pittsburgh.

In this section, we investigate whether principals' selection of teachers is likely to bias the validity results presented above. We compare the characteristics of teachers who did and did not participate in Phase 2. Large differences in teacher characteristics would suggest that correlations may not be representative and that bias could be important. Then we assess the robustness of our main correlation results to the selection of Phase 2 teachers by accounting for characteristics of teachers' students when calculating correlations.[23] By accounting directly for teacher characteristics, we reduce the influence of teacher selection on the estimated correlations. This approach relies on comparisons of more similar teachers and reduces the chance that observed correlations are driven by teachers who teach very different students. Basing the correlations on comparisons between only similar teachers could avoid spurious correlations from comparing teachers who were selected by different principals, such as in the example above. Although we cannot hope to eliminate all selection bias using this approach—because we cannot account for all factors related to selection—the results could suggest the bias is small if selection-adjusted correlations are not substantively different from those in Table IV.1.[24]

1. Differences in Teacher Characteristics Between Phase 2 Participants and Nonparticipants

Table IV.3 presents differences in teacher characteristics between teachers who did and did not participate in Phase 2. The top panel pertains to Pennsylvania teachers outside Pittsburgh, and the bottom panel pertains to Pittsburgh. The first pair of columns shows the mean and sample size for Phase 2 teachers, the second pair shows these statistics for non-Phase 2 teachers, and the final pair shows the raw difference and the difference adjusted for the characteristics of teachers' students. We restrict the teachers in the table to those with value-added estimates, since this is the sample of teachers used to calculate correlations.

Among non-Pittsburgh teachers, those who participated in Phase 2 are more likely to teach advantaged students, defined by average pre-test scores and student characteristics. For example, the average Phase 2 teacher's students score 15.3 PSSA points higher on the previous year's reading test. This difference is equivalent to moving a student at the median of the achievement distribution to the 53rd percentile. This difference is statistically significant—as are differences with respect to most characteristics in Table IV.3 for non-Pittsburgh teachers—and suggests that selection might lead to important bias in the results. In contrast, fewer of the characteristics show statistically significant differences between Pittsburgh teachers that participated and did not participate in Phase 2. Accounting for characteristics of teachers' students generally reduced differences between Phase 2 and non-Phase 2 teachers, as indicated by the smaller adjusted differences in the last column of

[23] We account for the average characteristics of teachers' students in 2011–2012. The characteristics include indicators for each grade taught, indicators for each subject taught, average pre-test scores in math and reading, and the fraction of students with each of the following characteristics: free or reduced-price lunch status, English-language-learner status, special education status, attended multiple schools during the year, male, black, Hispanic, Asian, and other race/ethnicity.

[24] One concern with this approach is that accounting for characteristics of teachers' students in the correlations could *over-control* for differences in teacher effectiveness across schools. In other words, teachers in one school might be more effective for reasons unrelated to the rubric, possibly, for example, because they have a particularly effective principal, and so accounting for this difference—via student characteristics that may be proxies for principal effectiveness—incorrectly results in higher correlations.

Table IV.3. For example, the adjusted difference in reading scores was 4.5 PSSA points, equivalent to moving a student at the median of the achievement distribution to the 51st percentile.

Table IV.3. Characteristics for Phase 2 and Other Teachers—Pittsburgh and Non-Pittsburgh Samples

	Phase 2		Not Phase 2		Phase 2 Minus Not Phase 2	
	Average	Teachers	Average	Teachers	Unadjusted	Adjusted
Teachers Outside Pittsburgh						
Characteristics of Teachers' Students						
Average prior math PSSA (mean-centered scaled score)	5.7	271	−13.6	24,840	19.3*	3.4
Average prior reading PSSA (mean-centered scaled score)	1.5	271	−13.9	24,840	15.3*	−4.5*
Fraction of students with free or reduced-price lunch status	0.41	271	0.44	24,840	−0.03	0.02*
Fraction of English-language learners	0.01	271	0.03	24,840	−0.01*	−0.01*
Fraction of special education students	0.16	271	0.21	24,840	−0.04*	−0.03*
Fraction Female Teachers	0.72	228	0.76	21,227	−0.04	−0.01
Total Experience						
Fraction five years or fewer	0.18	228	0.16	21,227	0.02	0.03
Fraction more than five years	0.82	228	0.84	21,227	−0.02	−0.03
Annual Salary ($)	56,600	228	61,800	21,200	−5,200*	−3,200*
Teachers in Pittsburgh						
Characteristics of Teachers' Students						
Average prior math PSSA (mean-centered scaled score)	−87.9	395	−80.9	82	−7.0	−5.2
Average prior reading PSSA (mean-centered scaled score)	−109.4	395	−104.7	82	−4.7	5.5
Fraction of students with free or reduced-price lunch status	0.78	395	0.77	82	0.00	0.01
Fraction of English-language learners	0.77	395	0.74	82	0.03*	0.01
Fraction of special education students	0.24	395	0.22	82	0.02	0.01
Fraction Female Teachers	0.77	387	0.75	75	0.02	0.00
Total Experience						
Fraction five years or fewer	0.19	387	0.08	75	0.11*	0.12*
Fraction more than five years	0.81	387	0.92	75	−0.11*	−0.12*
Annual Salary ($)	71,600	387	74,200	75	−2,600	−3,100

Source: Mathematica calculations based on Phase 2 classroom observation data in the 2011–2012 school year and characteristics of 2011–2012 teachers in Pennsylvania's longitudinal student database.

Notes: Adjusted differences account for indicators for each grade taught, indicators for each subject taught, average pre-test scores in math and reading, and the fraction of students with each of the following characteristics: free or reduced-price lunch status, English-language-learner status, special education status, attended multiple schools during the year, male, black, Hispanic, Asian, and other race/ethnicity. The characteristic that is averaged in the row is excluded when calculating the adjustment.

PSSA scores are mean-centered within grade based on all students who took the PSSA within each grade. The standard deviation for prior PSSA scores is 223 in math and 215 in reading.

The table includes Phase 2 teachers with value-added estimates.

* Statistically significant at the 5 percent level.

2. Differences in Effectiveness of Phase 2 Participants and Nonparticipants

The statistically significant differences in student characteristics noted in Table IV.3 raise a possible concern that the teachers who participated in Phase 2 were selected based on their effectiveness, but this does not appear to be the case. Differences in value-added estimates for

Phase 2 and non-Phase 2 teachers are in Table IV.4. The value-added estimates for Phase 2 teachers range from 5.6 PSSA points smaller to 8.5 points larger than the value-added estimates of non-Phase 2 teachers, depending on grade level and subject assignment. None of these differences are statistically significant, and the magnitude of the difference is not large. For example, 8.5 PSSA points is equivalent to moving a teacher at the median of the effectiveness distribution to the 57th percentile. These findings suggest that differences in value-added estimates between Phase 2 and non-Phase 2 teachers are not likely to generate substantial bias. The final column of Table IV.4 shows the difference in value-added estimates after accounting for the other teacher-level characteristics in the table. The magnitude of differences is typically smaller after adjusting for other characteristics. This suggests that any bias in the correlations can also be reduced by accounting for these characteristics.

Table IV.4. Value-Added Estimates for Phase 2 and Other Teachers—Non-Pittsburgh Sample

	Phase 2		Not Phase 2		Difference	
	Mean	Teachers	Mean	Teachers	Unadjusted	Adjusted
Teachers Outside Pittsburgh						
Value added, all teachers	−2.4	271	−2.8	24,840	0.5	0.0
Value added, homeroom teachers in grades 4-6	5.6	94	−1.4	11,305	7.0	4.3
Value added, departmentalized math teachers in grades 6-8	−6.4	44	−3.8	4,087	−2.7	−1.5
Value added, departmentalized ELA teachers in grades 6-8	−9.3	74	−3.8	6,094	−5.6	−3.8
Value added, departmentalized science teachers in grades 6-8	2.8	29	−1.5	1,491	4.3	0.8
Teachers in Pittsburgh						
Value added, all teachers	−13.9	395	−22.4	82	8.5	6.0

Source: Mathematica calculations based on Phase 2 classroom observation data in the 2011–2012 school year and value-added estimates using data from 2009–2010 through 2011–2012.

Notes: Adjusted differences account for grade and subject taught, average pre-test scores in math and reading, and the fraction of students with each of the following characteristics: free or reduced-price lunch status, English-language-learner status, special education status, attended multiple schools during the year, male, black, Hispanic, Asian, and other race/ethnicity.

No differences in the table are statistically significant. We exclude the subgroups for Pittsburgh because differences in value-added estimates are imprecise for these groups due to samples of 25 or fewer non-Phase 2 teachers. No differences for these subgroups were statistically significant

The table includes Phase 2 teachers with value-added estimates.

3. Correlations Between Rubric Scores and Value Added that Adjust for Selection

Accounting for selection did not substantively change correlations between value added and the rubric scores. Table IV.5 shows the correlations between rubric domains and value-added estimates with and without accounting for principal selection of teachers. The first column shows the unadjusted baseline correlations between rubric domains and value-added estimates for Phase 2 teachers in Pittsburgh; these are the same correlations as in Table IV.1. The second column shows adjusted correlations for Pittsburgh teachers that account for the characteristics of teachers' students. For Pittsburgh teachers, these correlations were smaller in magnitude, but the differences may not be substantive.[25] For example, the baseline correlation between value added and the PPR decreased from 0.22 to 0.20 after accounting for selection. The remaining columns show the same correlations for teachers outside Pittsburgh. In contrast to the Pittsburgh results, correlations for other teachers increased after adjusting for teacher characteristics. For example, the correlation between value added and the PPR increased from 0.22 to 0.24. As with the Pittsburgh results, the differences may not be substantive.

Table IV.5. Validity of Rubric Domains and Professional Practice Rating Accounting for Selection—All Districts

	Pittsburgh			Not Pittsburgh		
	Baseline Correlation	Selection-Adjusted Correlation	Teachers	Baseline Correlation	Selection-Adjusted Correlation	Teachers
Domain 1: Planning and Preparation	0.20*	0.17*	369	0.21*	0.23*	268
Domain 2: Classroom Environment	0.18*	0.17*	372	0.16*	0.17*	265
Domain 3: Instruction	0.27*	0.25*	389	0.24*	0.27*	271
Domain 4: Professional Responsibilities	0.16*	0.13*	368	0.11	0.11	265
Professional Practice Rating	0.22*	0.20*	395	0.22*	0.24*	271

Source: Mathematica calculations based on Phase 2 classroom observation data in the 2011–2012 school year and value-added estimates from the 2009–2010 through 2011–2012 school years.

Notes: All correlations in the table account for estimation error in value-added estimates (Jacob and Lefgren 2008). Selection adjusted correlations account for grade and subject taught, average pre-test scores in math and reading, and the fraction of students with each of the following characteristics: free or reduced-price lunch status, English-language-learner status, special education status, attended multiple schools during the year, male, black, Hispanic, Asian, and other race/ethnicity.

Baseline correlations are the same correlations reported in Table IV.1.

* Statistically significant at the 5 percent level.

Based on these findings, the nonrandom selection of Phase 2 teachers may not lead to large bias in the correlations. However, our methods to account for this selection also may not fully address this bias and so the impact of selection on the correlations could be larger than the differences reported in Table IV.5. Given how the data were collected, no analysis would warrant a conclusion that principal selection did or did not lead to bias. These results suggest that we might expect to see similar correlations had all teachers participated in Phase 2, though our analysis could account for selection based only on factors that are observable in the data.

D. Relationships Between the Rubric and Value Added Are Robust to VAMs Based on Different Cohorts of Students

Finally, we assess whether the correlations between rubric scores and value added depend on the student cohorts that are included in the VAM. Specifically, we estimate four VAMs: (1) a three-

[25] For Pittsburgh, where there are typically multiple Phase 2 teachers in each school, we also accounted for principals directly by including school fixed effects. The correlations are similar to those in the table, with one exception. The correlation for Domain 4 falls from 0.13 to 0.03 and is no longer statistically significant when school fixed effects are included.

cohort VAM that includes student achievement growth from the school year of the Phase 2 rubric observations and the previous two cohorts; (2) a three-cohort VAM that includes student achievement growth from the year prior to the school year of the Phase 2 rubric observations and the previous two cohorts; (3) a one-cohort VAM that includes student achievement growth only from the school year of the Phase 2 rubric observations; and (4) a one-cohort VAM that includes student achievement growth from only the year prior to the school year of the Phase 2 rubric observations.

Our results thus far have been based on the three-cohort VAM that combined the 2011–2012 cohort of students (that is, from the Phase 2 year) with the prior two cohorts. This VAM used the most recent three cohorts available for estimation, including student achievement from the school year of the Phase 2 rubric evaluations, and best approximates the sample that PDE might recommend for VAM estimates used in actual evaluations. However, including the contemporaneous cohort in a VAM used for the validation of the rubric could lead to correlations that *overstate* the validity of the rubric (Rockoff and Speroni 2011; Kane et al. 2011; Kane and Staiger 2012). This potential bias arises because the rubric score and the value-added estimate for a teacher are based on some of the same students. This will lead the correlations to overstate validity if some students have unexpectedly low growth in achievement during the year—for reasons unrelated to the effectiveness of the teacher—and the presence of these same students in a classroom results in a lower rubric score. For example, a principal—even with the best intentions to apply the rubric faithfully—might be influenced to give a teacher a lower score on an component because of several disruptive students in class and not because the teacher is unskilled. The teacher's score might have better reflected her or his true ability had the students in the classroom had fewer behavioral issues. If the VAM fails to account fully for the same characteristics of students in the classroom, then the parallel mis-measurement of the teacher's ability in both the rubric and the VAM could be responsible for any observed association between the scores.

We also estimated a VAM that combines the three cohorts of students prior to 2011–2012 so that the VAM no longer includes students in teachers' classrooms when the Phase 2 observations took place. This alternative approach reduces the type of bias described above. However, it may *understate* validity because there could be real differences in teacher effectiveness between the years examined by each measure, which would lower the degree to which the measures correlate. We estimate this VAM for all teachers, but because of limitations in the available data, this VAM includes only two cohorts of students for Pittsburgh teachers—the 2009–2010 and 2010–2011 school years—instead of the three cohorts included for teachers in other Pennsylvania districts.

Finally, we estimate two one-cohort VAMs that include only students from the 2010–2011 or 2011–2012 school year. While one-cohort VAMs yield less precise estimates than three-cohort VAMs, using only the most recent cohorts of students could provide the most accurate measure of teacher effectiveness in 2011–2012. As with the three-cohort VAMs, the one-cohort 2010–2011 VAM may understate validity, and the one-cohort 2011–2012 VAM may overstate validity.

As expected, correlations with rubric scores based on VAMs that include students from the 2011–2012 school year are larger than the same correlations based on VAMs that do not include the 2011–2012 school year. For example, the correlation for the PPR in our baseline three-cohort model that includes 2011–2012 was 0.24 (first column of Table IV.6). The same correlation for the three-cohort model that excludes 2011–2012 was 0.20 (second column of Table IV.6). The above discussion of potential bias in these results suggest that the correlation between the rubric scores and a teacher's actual effectiveness—as would be measured by a VAM with no bias—probably lies between these results. Results for the one-cohort VAMs were similar to those for the three-cohort

VAMs but with a wider range of values: the correlations with the PPR were 0.27 including 2011–2012 and 0.16 excluding 2011–2012 (last two columns of Table IV.6). The correlation for domain 4—professional responsibilities—is under 0.1 for both VAMs that exclude 2011–2012, and is not statistically significant for the three-cohort model. We report correlations for individual components with the each of three alternative VAMs in Appendix D.

Table IV.6. Validity of Rubric Domains and Professional Practice Rating by Value-Added Model—All Districts

	Three-Cohort Value-Added Models				One-Cohort Value-Added Models			
	2009–2010 Through 2011–2012		2008–2009 Through 2010–2011		2011–2012		2010–2011	
	Corr.	Teachers	Corr.	Teachers	Corr.	Teachers	Corr.	Teachers
Domain 1: Planning and Preparation	0.23*	637	0.18*	528	0.23*	578	0.14*	498
Domain 2: Classroom Environment	0.19*	637	0.17*	529	0.20*	579	0.14*	500
Domain 3: Instruction	0.28*	660	0.24*	550	0.29*	601	0.20*	520
Domain 4: Professional Responsibilities	0.17*	633	0.07	524	0.25*	576	0.09*	496
Professional Practice Rating	0.24*	666	0.20*	556	0.27*	607	0.16*	526

Source: Mathematica calculations based on Phase 2 classroom observation data in the 2011–2012 school year and value-added estimates using data from 2008–2009 through 2011–2012.

Notes: Because of data limitations, the three-cohort VAM that excludes 2011–2012 does not include students from the 2008–2009 school year for Pittsburgh teachers.

Correlations are adjusted for estimation error in value-added estimates (Jacob and Lefgren 2008).

* Statistically significant at the 5 percent level.

There are two concerns with interpreting these correlations as upper and lower bounds. First, as mentioned earlier, principals in Pittsburgh had received value-added information based on the 2010–2011 year for their schools and teachers, which they may have used in reaching conclusions about teachers' professional practices in 2011–2012. If so, it would tend to increase all correlations in Table IV.6. Second, the sample of teachers is not consistent across the models. A key reason the samples are different is that novice teachers in the 2011–2012 school year do not receive a value-added estimate from VAMs that exclude this school year. We estimated correlations based on a consistent sample of teachers across all VAMs—excluding novice teachers in 2011–2012. The results were substantively unchanged, which suggests that differences in the composition of teachers with estimates from different VAMs were not responsible for the estimates in Table IV.6 (Appendix Table D.4). Overall, the consistently positive correlations in this robustness check (with the sole exception of the professional responsibilities domain) support our main finding that there are positive and significant relationships between the rubric and value added that do not depend on the sample of students or teachers used to calculate the correlations.

V. CONCLUSIONS

As Pennsylvania nears its 2013–2014 date for rolling out a new teacher evaluation system statewide, its experiences from multiple pilot years can now guide implementation. Since 2010–2011, much has been learned through piloting the rubric about the importance of training, the rating scores that are likely to be obtained, and the ways in which rubric components relate to teachers' impacts on student learning.

In this Phase 2 report, we focused on assessing the rubric's internal consistency, score variation, and relationships with value added. First, we found that the rubric has an acceptable level of internal consistency, which means that teachers' ratings tended to be similar across the components in a domain. Our findings support a conclusion that the fairness of teachers' overall scores may not be compromised substantially by principals using different sets of components to rate teachers' professional practices in rubric domains. However, internal consistency generally increases as additional components are included in ratings, so principals should endeavor to use all components where evidence for a rating exists. Next, we found that most teachers were rated as either *proficient* or *distinguished* on each rubric component. Despite the narrow range of component scores, because summary-level rubric measures combine scores from multiple components, they have more potential to differentiate effectiveness among teachers. Even so, because value added is likely more precise, it may have the most potential to differentiate. Although principals appear to differentiate teacher performance using the rubric, they might not apply it consistently. Our findings reinforce the value of training on the rubric, even for principals who have previously been trained on it.

Finally, we found that teachers with higher rubric scores tended to make larger contributions to student achievement. The correlations are typically small—about 0 to 0.28—but they are comparable to those found by previous researchers, and most are statistically significant. Across the domains, the largest correlations tended to be found in domain 3 (instruction) and the smallest in domain 4 (professional responsibilities). Practices measured by components in the instruction domain may provide especially promising targets for professional development because of their stronger relationship with student achievement growth. The findings are also consistent with PDE's decision to give more weight to domain 3 and less to domain 4 in the calculation of teachers' PPRs. The correlations were robust to several alternative specifications and, among grade spans and subjects, were noticeably largest in middle-school science. Our findings support a conclusion that the rubric is measuring aspects of teachers' practices that are related to growth in student achievement on standardized assessments.

Despite the progress thus far through the pilot, there is still more to be learned about the rubric in the third phase and beyond. For example, no large-scale trial implementation has been conducted of the rating tool that will incorporate rubric ratings along with data on building-level achievement, student growth, and elective measures to obtain teachers' final ratings. Further, there have been no formal tests of the rubric's inter-rater and test-retest reliability in the context of actual teacher observations. Since some teachers will have participated in multiple rounds of the pilot by Phase 3, it would be possible to document the year-to-year consistency of their rubric ratings and value-added estimates, and to describe how useful the different measures that will constitute the overall evaluation system are for predicting teachers' future effectiveness. The larger, more representative sample that is expected for Phase 3 could also be used to understand better how principals select which components to use, and whether some principals' ratings of teachers are more strongly related than others to teachers' contributions to student achievement. Finally, replicating our current analyses using a broader, statewide sample would be best way to ensure that the findings presented in this report can be generalized beyond the teacher sample that was available during Phase 2.

This page has been left blank for double-sided copying.

REFERENCES

Aaronson, Daniel, Lisa Barrow, and William Sander. "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics,* vol. 25, no. 1, 2007, pp. 95–135.

Benjamin, Woan-Jue. "Development and Validation of Student Teaching Performance Assessment Based on Danielson's Framework for Teaching." Unpublished Manuscript. April 2002.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics,* vol. 119, no. 1, 2004, pp. 249–275.

Buonaccorsi, John P. *Measurement Error: Models, Methods, and Applications*. Boca Raton, FL: Chapman & Hall/CRC, 2010.

Chaplin, Duncan, Brian Gill, Allison Thompkins, and Hannah Miller. "Multiple Measures of Teacher Performance in the Pittsburgh Public Schools." Draft report. Washington, DC: U.S. Department of Education, Mid-Atlantic Regional Education Laboratory at ICF International, 2013.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood." Working paper #17699. Cambridge, MA: National Bureau of Economic Research, 2011.

Cortina, Jose M. "What Is Coefficient Alpha? An Examination of Theory and Applications." *Journal of Applied Psychology,* vol. 78, no. 1, 1993, pp. 98–104.

Cronbach, Lee J. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika,* vol. 16, no. 3, 1951, pp. 297–334.

de Vaus, David A. *Surveys in Social Research,* 5th edition. Crows Nest, Australia: Allen & Unwin, 2002.

Glazerman, Steven, Susanna Loeb, Dan Goldhaber, Douglas Staiger, Stephen Raudenbush, Grover Whitehurst. *Evaluating Teachers: The Important Role of Value-Added*. Washington, DC: Brown Center on Education Policy at Brookings, 2010.

Goldhaber, Dan, and Duncan Chaplin. "Assessing the Rothstein Test: Does It Really Show Teacher Value-Added Models Are Biased?" Washington, DC: Mathematica Policy Research, 2012.

Grossman, Pamela, Susanna Loeb, Julia Cohen, Karen Hammerness, James Wyckoff, Donald Boyd, and Hamilton Lankford. "Measure for Measure: The Relationship Between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value-Added Scores." Working paper #45. National Center for Analysis of Longitudinal Data in Education Research, 2010.

Hanushek, Eric A., John F. Kain, Daniel M. O'Brien, and Steven G. Rivkin. "The Market for Teacher Quality." Working paper #11154. Cambridge, MA: National Bureau of Economic Research, 2005.

Harris, Douglas N., and Tim R. Sass. "What Makes for a Good Teacher and Who Can Tell?" Unpublished manuscript, 2010.

Herrmann, Mariesa, Elias Walsh, Eric Isenberg, and Alex Resch. "Shrinkage of Value-Added Estimates and Characteristics of Students with Hard-to-Predict Achievement Levels." Working paper. Washington, DC: Mathematica Policy Research, 2013.

Ho, Andrew D., and Thomas J. Kane. *The Reliability of Classroom Observations by School Personnel.* Seattle, WA: Bill & Melinda Gates Foundation, 2013.

Hock, Heinrick, and Eric Isenberg. "Methods for Accounting for Co-Teaching in Value-Added Models." Working paper. Washington, DC: Mathematica Policy Research, 2012.

Jacob, Brian A., and Lars Lefgren. "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education." *Journal of Labor Economics,* vol. 26, no. 1, 2008, pp. 101–136.

Kane, Thomas J., and Douglas O. Staiger. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." Working paper #14607. Cambridge, MA: National Bureau of Economic Research, 2008.

Kane, Thomas J., and Douglas O. Staiger. *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains.* Seattle, WA: Bill & Melinda Gates Foundation, 2012.

Kane, Thomas J., Eric S. Taylor, John H. Tyler, and Amy L. Wooten. "Identifying Effective Classroom Practices Using Student Achievement Data." *Journal of Human Resources,* vol. 46, no. 3, 2011, pp. 587–613.

Kimball, Steven M., Brad White, Anthony T. Milanowski, and Geoffrey Borman. "Examining the Relationship Between Teacher Evaluation and Student Assessment Results in Washoe County." *Peabody Journal of Education,* vol. 79, no. 4, 2004, pp. 54–78.

Lane, Suzanne, and Christy Horner. "Pennsylvania Teacher and Principal Evaluation Pilot Report. Final Report Submitted to the Team Pennsylvania Foundation." Pittsburgh, PA: University of Pittsburgh, 2011.

Lipscomb, Stephen, Hanley Chiang, and Brian Gill. "Value-Added Estimates for Phase 1 of the Pennsylvania Teacher and Principal Evaluation Pilot." Cambridge, MA: Mathematica Policy Research, 2012.

McCaffrey, Daniel F., J. R. Lockwood, Daniel Koretz, Thomas A. Louis, and Laura Hamilton. "Models for Value-Added Modeling of Teacher Effects." *Journal of Educational and Behavioral Statistics,* vol. 29, no. 1, 2004, pp. 67–102.

Meyer, Robert H., "Value-Added Indicators of School Performance: A Primer." *Economics of Education Review,* vol. 16, no. 3, 1997, pp. 283–301.

Milanowski, Anthony. "The Relationship Between Teacher Performance Evaluation Scores and Student Achievement: Evidence from Cincinnati." *Peabody Journal of Education,* vol. 79, no. 4, 2004, pp. 33–53.

Morris, Carl N. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association,* vol. 78, no. 381, 1983, pp. 47–55.

Protik, Ali, Elias Walsh, Alexandra Resch, Eric Isenberg, and Emma Kopa. "Does Tracking of Students Bias Value-Added Estimates for Teachers?" Working paper. Washington, DC: Mathematica Policy Research, 2013.

Raudenbush, Stephen W. "What Are Value-Added Models Estimating and What Does This Imply for Statistical Practice?" *Journal of Educational and Behavioral Statistics,* vol. 29, no. 1, 2004, pp. 121–129.

Rockoff, Jonah E. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review,* vol. 94, no. 2, 2004, pp. 247–252.

Rockoff, Jonah E., and Cecilia Speroni. "Subjective and Objective Evaluations of Teacher Effectiveness: Evidence from New York City." *Labour Economics,* vol. 18, no. 5, 2011, pp. 687–696.

Rockoff, Jonah E., Douglas O. Staiger, Thomas J. Kane, and Eric S. Taylor. "Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools." Working paper #16240. Cambridge, MA: National Bureau of Economic Research, July 2010.

Sanders, William L. "Value-Added Assessment from Student Achievement Data—Opportunities and Hurdles." *Journal of Personnel Evaluation in Education,* vol. 14, no. 4, 2000, pp. 329–339.

Sawchuck, Stephen. "Teachers' Ratings Still High Despite New Measures: Changes to Evaluation Systems Yield Only Subtle Differences." *Education Week,* February 6, 2013.

Stacy, Brian, Cassandra Guarino, Mark Reckase, and Jeffrey Wooldridge. "Does the Precision and Stability of Value-Added Estimates of Teacher Performance Depend on the Types of Students They Serve?" Unpublished manuscript. October 2012.

Sturman, Michael C., Robin A. Cheramie, and Luke H. Cashen. "The Impact of Complexity and Performance Measurement on the Temporal Consistency, Stability, and Test-Retest Reliability of Employee Job Performance Ratings." *Journal of Applied Psychology,* vol. 90, no. 2, 2005, pp. 269–283.

Wasserman, John D., and Bruce A. Bracken. "Psychometric Characteristics of Assessment Procedures." In *Handbook of Psychology: Assessment Psychology,* vol. 10, edited by Irving B. Weiner, John R. Graham, and Jack A. Naglieri. Hoboken, NJ: John Wiley and Sons, 2003.

Weisberg, Daniel, Susan Sexton, Jennifer Mulhern, and David Keeling. *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness.* Brooklyn, NY: The New Teacher Project, 2009.

This page has been left blank for double-sided copying.

APPENDIX A

DATA SOURCES AND SAMPLE CHARACTERISTICS

In this appendix, we describe the data sources and characteristics of the teacher and student samples. In Section A, we cite the primary data sources. In Section B, we show sample characteristics for the classroom observation rubric data, and in Section C, sample characteristics for the students and teachers included in the value-added models (VAMs).

A. Data Sources

We use two primary types of data in this report: (1) scores from pilot administrations of the teacher evaluation rubric; and (2) statewide, student-level longitudinal data on academic outcomes and background characteristics.

1. Phase 2 Classroom Observation Rubric Scores

The Phase 2 rubric data includes classroom observation scores on 2,621 participating teachers. Of these teachers, 1,673 come from Pittsburgh Public Schools (Pittsburgh), and 948 come from 104 other Pennsylvania school districts. [26] The rubric data come from 2011–2012 except for most rubric scores for Pittsburgh teachers that participate in the district's Supported Growth Project (SGP). Teachers in the SGP focus their evaluations on one or more rubric components and retain their scores on other components from the previous year. Of the 2,621 total teachers in Phase 2, 666 teach in tested grades and subjects, and are also included in the validity analysis.

2. Statewide, Student-Level Longitudinal Data

Student-level longitudinal data were obtained to estimate teachers' effects on their students' academic outcomes. We obtained the math, reading, science, and writing scores of all Pennsylvania students in grades 3 through 8 on the Pennsylvania System of School Assessment (PSSA) and the PSSA-Modified from the Pennsylvania Department of Education's Bureau of Assessment and Accountability. These data cover school years 2006–2007 through 2011–2012. We also obtained student-level characteristics, course records, and school and teacher links from the Pennsylvania Information Management System (PIMS).[27] The PIMS data cover school years 2007–2008 through 2011–2012. Because of limitations with the PIMS data in Pittsburgh, we used Pittsburgh's own student-teacher links in 2009–2010 through 2011–2012, and used no student-teacher links for Pittsburgh from 2008–2009.

B. Descriptive Statistics on Phase 2 Classroom Observation Data

Tables A.1 through A.3 summarize the Phase 2 classroom observation data. For each rubric component, we report the number of teachers with a score, the percentage of these teachers in each

[26] Of the 1,673 Phase 2 teachers in Pittsburgh, 6 received rubric scores on only the two components used exclusively in Pittsburgh. Consequently, these teachers are not included in our analysis.

[27] We use the following PIMS templates: student (0320); course instructor (0410); course (0310); and staff (0330).

performance category, and the average rubric score. The tables summarize score data, respectively, for all Phase 2 teachers, for Phase 2 teachers in Pittsburgh, and for Phase 2 teachers outside Pittsburgh.

For use in the analysis, we calculate domain-level averages of these components, and a weighted average of these domain averages called the Professional Practice Rating (PPR). The domain averages are the average score across all the components in the domain for which the teacher received a Phase 2 score. The PPR is a weighted average of these domain-level scores. Domain-level scores for domains 1 and 4 are weighted 20 percent each, and for domains 2 and 3, 30 percent each. For teachers with no scores on any components in one or more domains, we scale up the weights for the domains in which they do have scores in proportion to sum to 100 percent. Consequently, all Phase 2 teachers have a PPR.

Table A.1. Summary of Phase 2 Classroom Observation Data—All Districts

Rubric Component	Teachers with Scores	Distinguished	Proficient	Needs Improvement	Failing	Average Score
1a: Demonstrating knowledge of content and pedagogy	1,782	17.1%	78.2%	4.6%	0.1%	2.12
1b: Demonstrating knowledge of students	2,336	19.9%	74.9%	5.2%	0.0%	2.15
1c: Setting instructional outcomes	2,312	12.6%	81.4%	6.0%	0.1%	2.06
1d: Demonstrating knowledge of resources	1,765	18.8%	75.5%	5.7%	0.0%	2.13
1e: Planning coherent instruction	2,039	16.5%	78.9%	4.5%	0.2%	2.12
1f: Designing ongoing formative assessments	1,780	8.0%	76.6%	15.3%	0.1%	1.93
2a: Creating a learning environment of respect and rapport	1,899	26.2%	67.2%	6.3%	0.3%	2.19
2b: Establishing a culture for learning	2,278	17.2%	74.2%	8.4%	0.1%	2.09
2c: Managing classroom procedures	1,914	17.8%	75.1%	7.1%	0.1%	2.11
2d: Managing student behavior	2,377	18.3%	71.7%	9.5%	0.4%	2.08
2e: Organizing physical space	1,764	14.3%	81.4%	4.1%	0.2%	2.10
3a: Communicating with students	1,744	17.4%	78.0%	4.5%	0.1%	2.13
3b: Using questioning and discussion techniques	2,223	8.1%	64.6%	27.2%	0.2%	1.80
3c: Engaging students in learning	2,519	14.8%	73.6%	11.5%	0.1%	2.03
3d: Using assessment to inform instruction	2,522	10.3%	72.2%	17.2%	0.2%	1.93
3e: Demonstrating flexibility and responsiveness	1,720	10.9%	83.9%	5.1%	0.1%	2.06
4a: Reflecting on teaching and student learning	2,405	14.7%	79.5%	5.7%	0.2%	2.09
4b: System for managing students’ data	2,299	11.7%	73.6%	14.4%	0.2%	1.97
4c: Communicating with families	2,252	18.0%	69.9%	11.9%	0.1%	2.06
4d: Participating in a professional community	1,769	19.4%	75.6%	4.7%	0.2%	2.14
4e: Growing and developing professionally	1,724	15.8%	79.5%	4.6%	0.1%	2.11
4f: Showing professionalism	1,580	20.3%	77.0%	2.5%	0.3%	2.17

Source: Mathematica calculations based on Phase 2 classroom observation data.

Table A.2. Summary of Phase 2 Classroom Observation Data—Pittsburgh Only

Rubric Component	Teachers with Scores	Distinguished	Proficient	Needs Improvement	Failing	Average Score
1a: Demonstrating knowledge of content and pedagogy	1,110	12.0%	82.0%	6.0%	0.0%	2.06
1b: Demonstrating knowledge of students	1,574	17.3%	77.3%	5.4%	0.0%	2.12
1c: Setting instructional outcomes	1,574	9.9%	83.3%	6.7%	0.1%	2.03
1d: Demonstrating knowledge of resources	1,101	9.7%	83.7%	6.5%	0.0%	2.03
1e: Planning coherent instruction	1,113	7.9%	86.3%	5.8%	0.0%	2.02
1f: Designing ongoing formative assessments	1,093	4.2%	74.8%	20.9%	0.1%	1.83
2a: Creating a learning environment of respect and rapport	1,117	19.4%	72.1%	8.1%	0.4%	2.11
2b: Establishing a culture for learning	1,578	15.5%	73.8%	10.5%	0.2%	2.05
2c: Managing classroom procedures	1,142	12.7%	78.9%	8.3%	0.1%	2.04
2d: Managing student behavior	1,579	15.3%	72.7%	11.5%	0.6%	2.03
2e: Organizing physical space	1,090	9.4%	86.0%	4.4%	0.3%	2.04
3a: Communicating with students	1,108	10.5%	84.7%	4.8%	0.1%	2.06
3b: Using questioning and discussion techniques	1,584	6.9%	60.4%	32.4%	0.3%	1.74
3c: Engaging students in learning	1,589	11.3%	74.2%	14.5%	0.1%	1.97
3d: Using assessment to inform instruction	1,594	8.3%	68.8%	22.8%	0.2%	1.85
3e: Demonstrating flexibility and responsiveness	1,101	6.1%	87.8%	6.0%	0.1%	2.00
4a: Reflecting on teaching and student learning	1,576	10.6%	82.9%	6.3%	0.1%	2.04
4b: System for managing students' data	1,575	9.7%	70.5%	19.5%	0.3%	1.90
4c: Communicating with families	1,574	17.8%	69.4%	12.6%	0.2%	2.05
4d: Participating in a professional community	1,105	12.0%	82.4%	5.3%	0.3%	2.06
4e: Growing and developing professionally	1,089	11.3%	82.7%	5.9%	0.1%	2.05
4f: Showing professionalism	965	13.0%	83.8%	2.9%	0.3%	2.09

Source: Mathematica calculations based on Phase 2 classroom observation data.

Table A.3. Summary of Phase 2 Classroom Observation Data—All Districts Except Pittsburgh

Rubric Component	Teachers with Scores	Distinguished	Proficient	Needs Improvement	Failing	Average Score
1a: Demonstrating knowledge of content and pedagogy	672	25.6%	72.0%	2.2%	0.1%	2.23
1b: Demonstrating knowledge of students	762	25.3%	69.9%	4.7%	0.0%	2.21
1c: Setting instructional outcomes	738	18.3%	77.2%	4.3%	0.1%	2.14
1d: Demonstrating knowledge of resources	664	33.9%	61.7%	4.4%	0.0%	2.30
1e: Planning coherent instruction	926	26.8%	69.9%	2.9%	0.4%	2.23
1f: Designing ongoing formative assessments	687	14.0%	79.5%	6.6%	0.0%	2.07
2a: Creating a learning environment of respect and rapport	782	35.9%	60.4%	3.6%	0.1%	2.32
2b: Establishing a culture for learning	700	21.1%	75.1%	3.7%	0.0%	2.17
2c: Managing classroom procedures	772	25.3%	69.6%	5.2%	0.0%	2.20
2d: Managing student behavior	798	24.4%	69.7%	5.8%	0.1%	2.18
2e: Organizing physical space	674	22.4%	74.0%	3.6%	0.0%	2.19
3a: Communicating with students	636	29.6%	66.5%	3.9%	0.0%	2.26
3b: Using questioning and discussion techniques	639	11.0%	75.0%	14.1%	0.0%	1.97
3c: Engaging students in learning	930	20.9%	72.7%	6.3%	0.1%	2.14
3d: Using assessment to inform instruction	928	13.9%	78.2%	7.7%	0.2%	2.06
3e: Demonstrating flexibility and responsiveness	619	19.4%	76.9%	3.6%	0.2%	2.16
4a: Reflecting on teaching and student learning	829	22.4%	72.9%	4.5%	0.2%	2.17
4b: System for managing students' data	724	16.3%	80.2%	3.5%	0.0%	2.13
4c: Communicating with families	678	18.4%	71.2%	10.3%	0.0%	2.08
4d: Participating in a professional community	664	31.8%	64.5%	3.8%	0.0%	2.28
4e: Growing and developing professionally	635	23.6%	74.0%	2.4%	0.0%	2.21
4f: Showing professionalism	615	31.9%	66.2%	1.8%	0.2%	2.30

Source: Mathematica calculations based on Phase 2 classroom observation data.

C. Sample Characteristics for Students and Teachers Included in the VAMs

In Table A.4, we report sample means for several student characteristics that are used in the VAMs. The data come from PIMS for 2011–2012 and pertain to the analysis samples for the math VAMs in grades 4 through 8. Sample characteristics for the VAMs based on other PSSAs are similar.

Table A.4. Descriptive Statistics on Student Characteristics by Math Value-Added Model, 2011–2012

Variable	Math 4	Math 5	Math 6	Math 7	Math 8
Average prior math PSSA (mean-centered scaled score)	8.0	7.1	9.0	10.6	14.3
Average prior reading PSSA (mean-centered scaled score)	6.6	6.8	8.4	9.7	13.3
Female (%)	48.8	49.0	49.0	49.0	48.9
White (%)	70.7	71.1	72.2	72.9	73.5
African American (%)	14.9	14.8	14.5	14.4	13.9
Hispanic (%)	8.8	8.9	8.3	8.2	7.9
Asian and Pacific Islander (%)	3.6	3.5	3.4	3.1	3.2
Multiracial or other race/ethnicity (%)	2.1	1.8	1.8	1.6	1.6
Free lunch eligibility (%)	38.2	37.1	35.5	34.3	32.7
Reduced-price lunch eligibility (%)	5.8	6.0	6.0	6.3	6.2
English-language learner (%)	2.5	2.4	2.2	2.1	2.0
Special education (%)	15.4	15.5	15.0	14.7	14.4
Grade repeater (%)	0.3	0.2	0.3	0.8	0.6
Number of students (1,000's)	107.9	110.5	114.7	121.2	119.2

Source: Mathematica calculations based on Pennsylvania student data.

Notes: The descriptive statistics in the table include students in the analysis sample for the one-cohort VAM for 2011–2012.

PSSA scores are mean-centered within grade based on all students who took the PSSA within each grade and year combination.

The grade-specific standard deviations for prior PSSA scores range from 221 to 225 in math and 213 to 217 in reading.

PSSA = Pennsylvania System of School Assessment.

In Table A.5, we describe potential and analysis samples for students in the teacher VAMs. By potential sample, we mean the number of students that have a nonmissing value of the outcome variable for a particular VAM. The analysis sample includes the subset of those students with nonmissing (or imputed) data on prior scores and student characteristics, and nonmissing data on teacher links. In addition, the VAM analysis samples were restricted to teachers who taught more than six student equivalents across all years included in the VAM. In this table, we report sample sizes for one-cohort VAMs based on students taught in 2011–2012 and for three-cohort VAMs based on students taught between 2009–2010 and 2011–2012.

Table A.5. Potential and Analysis Sample Sizes, by Outcome and Value-Added Model

	One Cohort (2011–2012)		Three Cohorts (2009–2010 to 2011–2012)	
Outcome	Students with Nonmissing Values of the Outcome Measure	Students in the Analysis Sample for Teacher VAMs	Students with Nonmissing Values of the Outcome Measure	Students in the Analysis Sample for Teacher VAMs
Math PSSA, grade 4	127,513	107,918	386,690	328,263
Math PSSA, grade 5	130,946	110,534	392,437	332,865
Math PSSA, grade 6	132,935	114,691	394,173	343,990
Math PSSA, grade 7	133,810	121,192	395,945	360,496
Math PSSA, grade 8	132,955	119,211	397,588	361,149
Reading PSSA, grade 4	127,169	107,998	385,659	327,324
Reading PSSA, grade 5	130,645	111,435	391,511	332,663
Reading PSSA, grade 6	132,667	116,385	393,310	347,106
Reading PSSA, grade 7	133,442	122,764	394,813	365,406
Reading PSSA, grade 8	132,652	120,948	396,558	366,568
Writing PSSA, grade 5	129,518	110,382	388,445	330,148
Writing PSSA, grade 8	131,076	119,341	392,677	363,210
Science PSSA, grade 4	127,071	101,831	385,470	310,739
Science PSSA, grade 8	131,616	119,990	394,423	361,374

Source: Mathematica calculations based on Pennsylvania student data.

Note: Sample sizes refer to unique student observations. Students are counted only once if they appear in a sample in multiple years. The analysis sample for an outcome measure is the sample that is used for estimating a VAM.

PSSA = Pennsylvania System of School Assessment; VAM = Value-added model.

In Table A.6, we report the number of teachers with VAM estimates by outcome, number of cohorts included in the VAM, and school years included in the VAM.

Table A.6. Number of Teachers with Value-Added Estimates, by Outcome and Value-Added Model

	One Cohort		Three Cohorts	
Outcome	2011–2012	2010–2011	2009–2010 to 2011–2012	2008–2009 to 2010–2011
Math PSSA, grade 4	4,846	5,063	5,026	5,257
Math PSSA, grade 5	4,722	4,852	4,959	5,101
Math PSSA, grade 6	3,222	3,396	3,640	3,818
Math PSSA, grade 7	2,124	2,148	2,723	2,788
Math PSSA, grade 8	2,084	2,104	2,696	2,761
Reading PSSA, grade 4	4,881	5,071	5,046	5,264
Reading PSSA, grade 5	4,839	4,946	5,063	5,203
Reading PSSA, grade 6	3,803	3,979	4,411	4,612
Reading PSSA, grade 7	2,774	2,834	3,655	3,798
Reading PSSA, grade 8	2,641	2,725	3,504	3,648
Writing PSSA, grade 5	4,832	4,944	5,101	5,248
Writing PSSA, grade 8	2,626	2,717	3,606	3,735
Science PSSA, grade 4	4,565	4,780	4,689	4,927
Science PSSA, grade 8	1,553	1,597	1,829	1,870
Teachers with at least one VAM estimate	23,942	24,667	25,610	26,411
Phase 2 teachers with at least one VAM estimate	607	526	666	556

Source: Mathematica calculations based on Pennsylvania student data.

Note: Teachers are included in multiple VAMs if they have students in multiple grades or subjects. The number of teachers with estimates excludes teachers whose estimates were based on fewer than 10 student equivalents across all grades and subjects they teach and teachers who did not teach any students in the most recent year included in the VAM (2010–2011 or 2011–2012).

PSSA = Pennsylvania System of School Assessment; VAM = Value-added model.

This page has been left blank for double-sided copying.

APPENDIX B

TECHNICAL FEATURES OF THE TEACHER VAMS

In this appendix, we provide a technical description of the teacher value-added models (VAMs). In Section A, we describe the empirical specification. In Section B, we describe the two-step method used to generate the teacher effectiveness estimates. In Section C, we list the prior-year achievement measures and other control variables. In Section D, we discuss how teachers' value-added estimates for different grades and subjects are combined to form teacher-level value-added scores.

A. The Empirical Model

The VAMs estimated in this report provide measures of teachers' contributions to student learning in 4th- through 8th-grade math and reading, 5th- and 8th-grade writing, and 4th- and 8th-grade science. We use Pennsylvania System of School Assessment (PSSA) and PSSA-Modified (PSSA-M) scores in these grades and subjects as outcomes, and students' own prior PSSA scores as baselines. The following regression equation, estimated separately for each grade-subject combination, describes the teacher VAMs:

$$(1) \qquad A_{itcy} = \beta' P_{i(y-1)} + \gamma' X_{iy} + \theta' C_{itcy} + \delta' T_{ity} + \varphi' Y_{y} + e_{itcy}$$

In the model, A_{itcy} is an assessment score for student i, taught by teacher t in class c, in year y. For example, A_{itcy} could be a 5th-grade PSSA math assessment. The sample would comprise student-teacher-class-year combinations across the state over a set period of years in which the student took a particular assessment and was taught by a particular teacher in the subject of the assessment. The vector $\mathbf{P}_{i(y\text{-}1)}$ includes school-year-specific variables for student i's prior-year PSSA scores. We include prior-year math and reading scores in all VAMs, and prior-year science and writing scores in VAMs where those scores would be available in the prior year. Including prior-year scores in two or more subjects captures a broader range of prior inputs than if only a same-subject prior-year score were used. For most students, prior-year scores come from the previous grade. However, prior scores for grade repeaters come from the same grade as the outcome variable. The vector $\mathbf{P}_{i(y\text{-}1)}$ therefore also includes a set of variables containing grade repeaters' same-grade PSSA scores from the previous year. Finally, the vector includes a variable for the test score from the prior-prior year in the same subject as the outcome for students in grades 5 through 8.

The vector $\mathbf{X}_{iy}$ is a set of variables for observed student characteristics. The vector $\mathbf{C}_{itcy}$ is a set of variables for the characteristics of student i's classroom peers. The vector $\mathbf{Y}_{y}$ includes year indicators for the school years in the VAM.[28] The coefficients in $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\theta}$ are the estimated relationships between students' assessment scores and each respective student characteristic, controlling for the other factors in the model. The variable e_{itcy} is the error term.[29]

[28] These indicators are excluded if the VAM includes only one cohort of student growth data.

[29] We use a standard cluster-robust variance estimator to obtain standard errors that adjust for clustering of observations by student and to account for heteroskedasticity.

The vector $\mathbf{T}_{ity}$ includes a teacher variable for each teacher in the VAM that is equal to one for students taught by the teacher, and zero otherwise. Students taught by multiple teachers are included in the model on multiple rows, once for each teacher, and each student-teacher-course-year observation has exactly one non-zero element in $\mathbf{T}_{ity}$. We use a weighted least squares regression to accurately attribute the exposure of students to teachers during the school year. This approach gives less weight to students in calculating a teacher's value added when students are also taught by another teacher in the same subject, grade, and year. A student contributes a total of 100 percent of his or her dosage to one or more teachers. A student's dosage is split between teachers in the event a student appears to take multiple courses in the same subject or to have multiple teachers in the same classroom (Hock and Isenberg 2012).

The vector $\boldsymbol{\delta}$ is a set of coefficients to be estimated, one for each teacher in the VAM. Each coefficient in $\boldsymbol{\delta}$ identifies a teacher's contribution to student learning—the extent to which the actual achievement of students tends to be above or below what is predicted for an average teacher. The average value-added score is set equal to zero, but this does not mean that student learning is zero for the teacher with the average value-added score. Rather, a positive value-added estimate represents above-average teacher performance and a negative estimate represents below-average performance. The reference point for determining the average teacher contribution depends on the sample of teachers in the model. Since the model includes students and teachers across the state, the value-added estimates are calculated relative to the contribution of the average teacher in Pennsylvania in the grade, subject, and school years covered by the VAM. Teachers' final value-added scores are based on a weighted average of these coefficient estimates (see Section D below).

In this report, we estimate VAMs that base teachers' value-added estimates on up to three years of student growth data—that is, the number of current and prior student cohorts who contribute to the estimate. Multiyear estimates are less prone to random and systematic fluctuations that stem from being assigned a few students with unusually high or low learning growth. They can therefore detect performance differences with greater validity and reliability. However, they are less reflective of immediate past performance than single-year estimates, as a result of combining annual value-added scores over multiple years. We therefore estimate single-year VAMs as well.

B. Two-Step Estimation Method

The VAMs rely on students' own prior achievement scores as indicators of their academic abilities before entering a teacher's classroom. Standardized tests are imperfect measures of students' true abilities. The measurement error introduced by using prior assessment scores as ability measures causes standard regression techniques to produce biased estimates of teacher effectiveness. We correct for the measurement error issue by incorporating the test/retest reliability of the PSSA tests into the regression models directly. This approach, called an errors-in-variables (EIV) regression, eliminates bias due to the known amount of measurement error in students' prior-year tests (Buonaccorsi 2010). In terms of equation (1), EIV provides a better estimate of $\boldsymbol{\beta}$ than would be obtained by ordinary regression.

Because of a technical limitation of the EIV approach, the VAMs must be estimated in two regression steps.[30] The two-step estimation method also allows us to address a separate technical issue that otherwise would prevent peer characteristics from being included in one-year VAMs in elementary grades. Elementary school teachers, unlike teachers in higher grades, typically teach the same students throughout the day in self-contained classrooms. When only one year of teaching data are included in a VAM, it is not possible to separate their contributions to student growth from the effects from students' classroom peers—thus leading to biased estimates of teacher effectiveness. Our solution, described below, it to estimate the relationships between peer characteristics and student performance in the first regression step based on three cohorts of students and then apply those relationships in the second stage based on just the one cohort of students. In terms of equation (1), this approach provides for estimates of $\boldsymbol{\theta}$ that otherwise might not be possible.

Specifically, we first estimate equation (1) separately for each grade-subject combination with the EIV correction for measurement error in the prior-year test scores, based on grade-specific reliability data for the PSSA published by the Pennsylvania Department of Education.[31] For all VAMs, the first stage regressions pool students and teachers from three school years. Using multiple years of data allows us to estimate the effects of classroom characteristics based on variation in peers across multiple classrooms for each teacher, instead of relying on variation between teachers and between schools. The latter sources of variation could lead to spurious associations between peer characteristics and student performance. We instead leverage variation across classrooms taught by individual teachers to identify the contribution of classroom composition to student achievement.[32]

Based on the results of estimating equation (1), we calculate an *adjusted* post-test for each grade, year, and subject that nets out the contribution of the measures of classroom composition and all prior test scores:

$$(2) \qquad \hat{A}_{itcy} = A_{itcy} - \beta' P_{i(y-1)} - \theta' C_{itcy} .$$

The vector $\hat{A}_{itcy}$ represents the student post-test outcome, net of the estimated contribution of classroom composition and prior test scores. We use the adjusted post-test in place of the actual post-test to estimate single-year or multiyear VAMs for the year or years of interest. To do this, we estimate equation (3) below separately for each grade-subject combination including only data from the year(s) of interest:

[30] The EIV model does not allow for standard errors to be clustered. Standard errors in our models must be clustered at the student level, because students can have multiple rows in the estimation file according to equation (1).

[31] The EIV correction is applied to the baseline scores in the second column of Table B.1 (see below). We do not apply the correction to the prior-prior year test scores or to the test scores of grade repeaters. There are higher fractions of imputed scores for prior-prior year tests than for other baselines, and relatively few students repeat grades between grades 4 and 8. The assumptions implicit in the EIV method about measurement error may not hold for these tests.

[32] To avoid potential bias from the sorting of teachers and students across schools, when estimating the contribution of classroom composition, we will not pool classrooms across schools for teachers who transfer between schools. Instead, for purposes of estimating the first-stage regression, we will treat teachers who transfer as being a separate teacher for each school in which he or she taught. The teacher's final value-added estimate is based on the teacher's students across all schools.

(3) $$\hat{A}_{itcy} = \gamma' X_{iy} + \delta' T_{ity} + \varphi' Y_{y} + e_{itcy}$$

We obtain standard errors that are consistent in the presence of both heteroskedasticity and clustering at the student level (which cannot be done directly in equation (1) using the EIV method), because the regression includes multiple observations for individual students.

This multistep method will likely underestimate the standard error of $\hat{\boldsymbol{\delta}}$ slightly, which means that the precision of teacher effectiveness estimates may be overstated, because the adjusted gain in equation (2) relies on the estimated values of $\boldsymbol{\beta}$, and $\boldsymbol{\theta}$. The magnitude of the underestimation is related to the precision of the coefficients used to calculate the adjusted gain, but may be ignorable.[33] With the large within-grade sample sizes over multiple years, the coefficient estimates on pre-test scores and peer-average characteristics are likely to be estimated precisely. Thus we believe that using the EIV approach to improve the accuracy of teacher effectiveness measures is worth the tradeoff with accurately measuring precision.

C. Prior-Year Achievement Measures and Control Variables

The VAMs cover each subject-grade combination in which the PSSA is given to students between grades 4 and 8 in math, reading, science, and writing. We account for prior-year student achievement on the PSSA math and reading assessments in all VAMs, and for 5th- through 8th-grade students, we account for their prior-year scores in the same subject as the outcome from the prior-prior year.[34] By including separate variables for PSSAs in each subject-grade-year combination, we allow the relationships between each prior-year test and achievement to vary across grade-year combinations. Students who repeat a grade are included in the VAM.[35] For these students, we include additional PSSA variables in math and reading, and separate indicators for repeating a grade in each year included in the VAM.[36]

Students do not take the science and writing PSSAs in consecutive grades, so for these outcomes we cannot account for prior-year scores in the same subject. The lack of a same-subject prior-year test is not necessarily a critical concern for estimating VAMs. Fundamentally, the VAM still determines whether student scores on an outcome (for example, 4th-grade PSSA science) are higher or lower than predicted for students with the same prior achievement scores (in this case measured by prior math and reading scores) and other background characteristics. We would expect, however, that VAMs with prior-year scores in the same subject will generally produce estimates that are more precise than those of VAMs that must rely on prior-year scores in other subjects.

[33] For example, because $\boldsymbol{\beta}$ is estimated, the error term in equation (3) is clustered within grades. This form of clustering typically results in estimated standard errors that are too small, because the second-step regression does not account for a common source of variability affecting all students in a grade. In view of the small number of grades, standard techniques of correcting for clustering will not effectively correct the standard errors (Bertrand et al. 2004).

[34] We use math scores as the same-subject baseline for science VAMs and reading scores as the same-subject baseline for writing VAMs.

[35] Students with very rare grade progressions—for example, students who appear to progress into a lower grade—are excluded from the VAMs.

[36] We do not include prior-prior year scores for students who repeated a grade.

Table B.1 summarizes the assessments used as outcome and baseline variables in the VAMs for students who do not repeat a grade. We require that students have at least one prior-year test score to be included in a VAM. We impute a small fraction of scores (less than 1 percent) for students who are missing one or more of the prior-year test scores. We also impute less than 7 percent of scores on the prior-prior year tests. The imputations are based on the relationships with other prior-year scores and observed characteristics of students who have nonmissing scores.

Table B.1. PSSAs Used as Outcomes and Baselines in the Teacher Value-Added Models

Outcomes		Prior-Year Baselines		Prior-Prior Year Baselines	
Subject	Grade	Subject	Grade	Subject	Grade
Math	4	Math, Reading	3	NA	n/a
Reading	4	Math, Reading	3	NA	n/a
Science	4	Math, Reading	3	NA	n/a
Math	5	Math, Reading, Science	4	Math	3
Reading	5	Math, Reading, Science	4	Reading	3
Writing	5	Math, Reading, Science	4	Reading	3
Math	6	Math, Reading, Writing	5	Math	4
Reading	6	Math, Reading, Writing	5	Reading	4
Math	7	Math, Reading	6	Math	5
Reading	7	Math, Reading	6	Reading	5
Math	8	Math, Reading	7	Math	6
Reading	8	Math, Reading	7	Reading	6
Science	8	Math, Reading	7	Math	6
Writing	8	Math, Reading	7	Reading	6

Note: Baseline scores for grade repeaters are their prior-year scores in the same grade as the outcome variable. We do not include prior-prior year scores for grade repeaters.

To help isolate the effect of teachers on student achievement, the VAMs also include control variables for observable student and peer background characteristics. Table B.2 lists these variables, which enter equation (1) through the vectors X_{iy} and C_{itcy}. The factors that are included are thought to be correlated with student performance and outside the control of teachers. A standard list of controls would include measures related to students' socioeconomic status (for example, parent educational attainment, family income, or proxies such as eligibility for free or reduced-price meals); family structure (for example, living in a single-parent household); or eligibility for programs such as special education.

There is usually a discrepancy between the variables that would ideally be included and the variables that are available in the data system. Researchers and policymakers are left with a choice between estimating a model that could systematically over- or under-estimate teacher contributions because of less-than-complete controls, and attempting to compensate at least partially for the omitted variables by including other measures that are available in the data. Most data systems collect only limited information on student background characteristics, typically basic demographic variables such as gender, race/ethnicity, meals program eligibility, disability status, and English-language-learner (ELL) status. Most researchers and policymakers opt to include whatever information is available while acknowledging that a different set of variables would be preferable, which is the approach we took in this report as well.

Table B.2. Student and Peer Characteristics Included in Value-Added Models for This Report

Control Variable	Definition
Free meals	Free meals eligibility {0,1}
Reduced-price meals	Reduced-price meals eligibility {0,1}
English-language learner (ELL)	ELL in outcome year {0,1}
Specific learning disability (SLD)	Designation of SLD under IDEA {0,1}
Speech or language impairment (SLI)	Designation of SLI under IDEA {0,1}
Emotional disturbance (ED)	Designation of ED under IDEA {0,1}
Intellectual disability (ID)	Designation of ID under IDEA {0,1}
Autism (AUT)	Designation of AUT under IDEA {0,1}
Physical/sensory impairment (PSi)	Designation of hearing impairment, visual impairment, deaf-blindness, or orthopedic impairment under IDEA {0,1}
Other impairment	Designation of other health impairment, multiple disabilities, developmental delay, or traumatic brain injury under IDEA {0,1}
Mobility	Attended multiple schools during school year {0,1}
Grade repeater	Repetition of the current grade (separate indicators by year in multiyear VAMs) {0,1}
Behind grade	More than 1.5 years older than expected for grade {0,1}
Age	Student age in years as of September 1
PSSA-modified (outcome)	Outcome is a PSSA-M score (PSSA outcomes only) {0,1}
PSSA-modified (prior-year math)	Prior-year math score is a PSSA-M score (prior-year PSSAs only) {0,1}
PSSA-modified (prior-year reading)	Prior-year reading score is a PSSA-M score (prior-year PSSAs only) {0,1}
Gender	Female {0,1}
Race/ethnicity	Indicators for African American, Hispanic, Asian Pacific Islander, or other race/ethnicity {0,1}
Classroom size	Number of students in the classroom
Classroom size interactions with student-level characteristics	Separate interaction terms between classroom size and the following student-level characteristics: ED, ID, AUT, PSI, any of the other three listed special education categories, free meals, and ELL
Peer-level characteristics	Separate peer-level averages for free meals, reduced-price meals, ELL, any special education category, gender, race/ethnicity categories, and prior math and reading test scores Also the peer-level standard deviation of prior math and reading test scores

Note: Peers are defined as a student's classmates in a particular classroom.

IDEA = Individuals with Disabilities Education Act; VAM = Value-added model; PSSA = Pennsylvania System of School Assessment; PSSA-M = PSSA-Modified.

As in Lipscomb et al (2012), we include controls for gender and race/ethnicity not to set different standards for students but rather as an empirical acknowledgement that these variables explain a statistically significant portion of the variation in student performance even after accounting for prior student achievement and all the other variables in Table B.2. To the extent that gender and race/ethnicity represent unobserved factors that differ across students and are outside the control of teachers, the VAM estimates would systematically penalize or reward certain teachers if these controls were omitted.

Finally, we include several classroom-level variables that account for peer influences on achievement. These measures are intended to account for various inputs that are largely beyond the control of teachers but affect their overall workload.[37] For example, accounting for average classroom achievement allows for the possibility that students perform better when they have higher-performing peers, and accounting for the standard deviation of classroom achievement allows the distribution of prior-year achievement in a classroom to affect student performance. These measures may also serve as an additional method of accounting for measurement error in individual student pre-test scores (to the extent that students' true achievement levels are related to their peers' test scores) (Protik et al. 2013).

D. Obtaining Teacher-Level Value-Added Scores

1. Combining Teacher Effectiveness Estimates Across Grades and Subjects

To obtain an overall value-added measure for each teacher, we combine teachers' value-added estimates for their grades and subjects. The composite measure, used in Chapters III and IV, represents the average contribution of teachers to their students' achievement growth across grades and subjects. To calculate the composite value-added measure, we standardize teachers' estimates to have the same variance across grades and subjects before combining all a teacher's estimates into a single composite score. A teacher's composite is obtained as an average of that teacher's grade- and subject-specific estimates. The average is weighted based on the number of the teacher's students who contribute to the VAM for the grade-subject combination. We also calculate the precision of teachers' composite measures based on the precision of their grade- and subject-specific estimates and the covariance between their estimates across subjects.[38]

The first step in calculating the composite is standardizing estimates to have the same scale across grades and subjects. Since the variability of value-added estimates may differ across grades and subjects, estimates of effectiveness for teachers in different grades and subjects may not be comparable. The main concern is that properties of the assessments used in the model—rather than teacher effectiveness—may drive discrepancies in the distribution of value-added estimates across grades and subjects. For example, larger student test score gains in a grade might reflect either additional student learning or a more sensitive test instrument. As a result of these differences, value-added estimates in some grades or subjects could influence the composite more than others even if they are given equal weight in an average. This counter-intuitive result could happen if, for instance, there is much more variation in math value-added estimates than reading value-added estimates. In this example, a teacher who is at the 85th percentile in math would appear to improve student achievement in math more than the 85th percentile reading teacher improves achievement in

[37] We do not include any measures related to educators' own characteristics (for example, years of experience) that might affect their effectiveness relative to that of other educators.

[38] We calculate the standard error of the combined estimate as the square root of the weighted sum of variances and covariances, divided by the total student equivalents taught by the teacher across all VAMs. The weights in the sum are the squared student equivalents for the specific VAM. We approximate each covariance as the correlation between value-added scores in the two subjects (within a grade), multiplied by the standard errors of a teacher's estimates in the subjects. We account for covariances only between subjects, and not between grades. This choice reflects the likelihood that teachers do not typically share many of the same students across the different grades they teach, whereas many teachers are responsible for instructing the same students in multiple subjects.

reading. Consequently, relative to reading, effective teaching in math would appear to "count more" in calculating a composite that includes both subjects. However, this difference may not reflect actual differences in math and reading teacher effectiveness.

Because we do not want to penalize or reward teachers simply for teaching a subject or grade with unusual test properties, we translate estimates within each subject-grade combination so that each set of estimates is expressed in a common metric of "generalized" PSSA points. Essentially we assume that the distribution of underlying teacher effectiveness is the same across grades and subjects, though we do not have a priori knowledge that this is truly the case. To translate the estimates, we calculate standardized scores (called *z-scores*) within each subject-grade combination by subtracting the average estimate from individual teachers' estimates and dividing by the standard deviation of estimates.[39] We then multiply the z-scores by a common factor based on the average variance across grade-subject combinations.[40] The units of the resulting value-added estimates reflect PSSA points in a "typical" grade and subject, and can therefore be compared and combined across grades and subjects. With the teacher effectiveness estimates expressed on the same scale, we then combine grade- and subject-specific estimates to obtain a set of final teacher-level estimates.

2. Empirical Bayes Shrinkage

Imprecision in value-added estimates can lead to very high or very low effectiveness measures for some teachers by chance. In the context of a high-stakes evaluation system, imprecision can lead to misclassification of teachers. Researchers often apply an empirical Bayes (EB) shrinkage procedure to reduce this chance. Whereas this approach is appropriate in most contexts, applying the EB procedure can also lead to bias in the results of analysis based on value-added estimates. This bias can often be avoided by addressing imprecision in value-added estimates by direct means in the analysis. For example, when calculating correlations between value-added estimates and rubric scores we use *pre-shrinkage* value-added estimates—estimates that are not adjusted using the EB procedure—and adjust the correlations for imprecision in value added using the method in Jacob and Lefgren (2008). In most of our analyses, we use the pre-shrinkage value-added estimates, because precision can be addressed by direct means in the analysis. However, when this is not possible, we use the EB *post-shrinkage* value-added estimates.

[39] A standard deviation measures score variation—what we can see graphically by whether the distribution of scores tends to be spread out or grouped tightly together. The standard deviations we use to calculate z-scores are adjusted for sampling error in the value-added estimates to avoid overstating the contributions of teachers to student achievement on the PSSA in some grades and subjects. The resulting estimates within each grade-subject combination will not have identical standardized variances, as is typical for z-scores. Rather, estimates in grades or subjects that are less precise will have lower standardized variances, reflecting the fact that these estimates carry less information about the teachers' contributions to student achievement. Consequently, estimates that are more precise receive more weight in the composite.

[40] We calculate this factor as the square root of the average variance across grade-subject combinations (weighted by the number of students taught by each teacher) within each subject-year combination. This factor reflects the standard deviation of teacher value added on the PSSA in a typical grade and subject, because students' PSSA scores were also standardized using a similar method. The variances are adjusted for sampling error in the value-added estimates to avoid overstating the contributions of teachers to student achievement on the PSSA.

To calculate EB estimates, we apply the EB procedure outlined in Morris (1983) to the composite teacher-level estimates.[41] Using the EB procedure, we compute a weighted average of an estimate for the average teacher (based on all students in the model) and the initial estimate that uses each teacher's own students. For teachers with relatively imprecise initial estimates based on their own students, the EB method effectively produces an estimate based more on the average teacher. For teachers with more precise initial estimates based on their own students, the EB method puts less weight on the value for the average teacher and more weight on the value obtained from the teacher's own students.

Shrinkage adjustments account for the fact that estimates with greater precision carry greater strength of information about teachers' true performance levels. The adjusted estimate is approximately equal to a precision-weighted average of the teacher's initial estimated effect and the overall mean of all estimated teacher effects, with more precise initial estimates receiving greater weight.[42] In essence, teachers are assumed to be average in performance until evidence justifies a different conclusion. By applying a greater degree of shrinkage to less precisely estimated teacher measures, the procedure reduces the likelihood that the estimate of effectiveness for a teacher falls at either extreme of the distribution by chance. We calculate standard errors for the EB estimates using the formulas provided by Morris (1983). As a final step, we remove estimates for any teachers with fewer than 10 student equivalents and then re-center the EB estimates on zero. We also apply this final step to the pre-shrinkage estimates we use in most of our analyses.

[41] In addition to the number of students taught by the teacher, student characteristics can also affect precision. Some students—such as special education students—have harder-to-predict test scores, which leads to more imprecise estimates for their teachers (Stacy et al. 2012; Herrmann et al. 2013).

[42] In Morris (1983), because of a correction for bias, the EB estimate does not exactly equal the precision-weighted average of the two values. This adjustment increases the weight on the overall mean by (K – 3)/(K – 1), where K is the number of teachers. We incorporate this correction in our shrinkage procedure.

This page has been left blank for double-sided copying.

APPENDIX C

TECHNICAL RESULTS FROM VALUE-ADDED ANALYSES

In this appendix, we provide technical results from the value-added models (VAMs) for estimating teacher effectiveness. In Table C.1, we report the amount of dispersion in the value-added estimates, the average standard error of an estimate, and the proportion of estimates that can be statistically distinguished from the average estimate. We report these separately for the one- and three-cohort VAMs that include the 2011–2012 school year. In Table C.2, we report the same information for the one- and three-cohort VAMs that do not include the 2011–2012 school year.

Table C.1. Sample Characteristics of Value-Added Models That End in the 2011–2012 School Year

	One-Cohort Value-Added Estimates 2011–2012			Three-Cohort Value-Added Estimates 2009–2010 to 2011–2012		
Outcome	84th Minus 50th Percentile of “Underlying” Value Added (In *z*-score units)	Average Standard Error (In *z*-score units)	Percentage of Estimates Statistically Distinguishable from the Average	84th Minus 50th Percentile of “Underlying” Value Added (In *z*-score units)	Average Standard Error (In *z*-score units)	Percentage of Estimates Statistically Distinguishable from the Average
Math PSSA, grade 4	0.25	0.11	42.3	0.24	0.08	52.0
Math PSSA, grade 5	0.24	0.09	45.7	0.23	0.07	53.4
Math PSSA, grade 6	0.23	0.08	51.8	0.23	0.07	57.5
Math PSSA, grade 7	0.20	0.07	56.3	0.23	0.06	59.2
Math PSSA, grade 8	0.19	0.07	55.2	0.22	0.07	58.2
Reading PSSA, grade 4	0.21	0.11	34.5	0.21	0.08	43.3
Reading PSSA, grade 5	0.18	0.10	31.9	0.18	0.08	41.8
Reading PSSA, grade 6	0.16	0.09	33.0	0.18	0.08	40.7
Reading PSSA, grade 7	0.15	0.07	39.5	0.19	0.08	45.2
Reading PSSA, grade 8	0.12	0.07	35.0	0.18	0.08	40.3
Writing PSSA, grade 5	0.36	0.13	51.0	0.35	0.10	61.1
Writing PSSA, grade 8	0.30	0.09	58.7	0.33	0.10	59.9
Science PSSA, grade 4	0.27	0.11	44.8	0.27	0.08	54.7
Science PSSA, grade 8	0.18	0.07	50.7	0.23	0.07	57.1
Combined teacher-level estimates (pre-shrinkage)	0.21	0.09	45.1	0.20	0.06	53.5
Combined teacher-level estimates (post-shrinkage)	0.21	0.08	41.4	0.20	0.06	51.3

Source: Mathematica calculations based on Pennsylvania student data.

Notes: The VAMs are based on statewide samples of teachers and students. Teachers’ VAM estimates are based on students in their classrooms at any time during the specified analysis periods.

One z-score unit is equal to one standard deviation of student outcomes. One standard deviation of student outcomes is equal to 227 PSSA points in math, 220 points in reading, 276 points in writing, and 190 points in science.

The 84th minus 50th percentile of underlying VAM estimates is the estimated difference in “underlying” value added for the teachers at these percentiles (that is, perfect measures of value added that do not have any estimation error). This is calculated as the standard deviation of value-added estimates with an adjustment for the amount of estimation error using the method in Morris (1983).

All estimates for individual subject-grade combinations are pre-shrinkage.

PSSA = Pennsylvania System of School Assessment; VAM = value-added model.

Table C.2. Sample Characteristics of Value-Added Models That End in the 2010–2011 School Year

	One-Cohort Value-Added Estimates 2010–2011			Three-Cohort Value-Added Estimates 2008–2009 to 2010–2011		
Outcome	84th Minus 50th Percentile of "Underlying" Value Added (In *z*-score units)	Average Standard Error (In *z*-score units)	Percentage of Estimates Statistically Distinguishable from the Average	84th Minus 50th Percentile of "Underlying" Value Added (In *z*-score units)	Average Standard Error (In *z*-score units)	Percentage of Estimates Statistically Distinguishable from the Average
Math PSSA, grade 4	0.27	0.12	40.7	0.26	0.09	50.9
Math PSSA, grade 5	0.24	0.10	44.5	0.24	0.07	53.2
Math PSSA, grade 6	0.24	0.08	51.0	0.24	0.07	59.5
Math PSSA, grade 7	0.20	0.07	54.7	0.23	0.07	58.4
Math PSSA, grade 8	0.17	0.07	47.1	0.22	0.07	53.8
Reading PSSA, grade 4	0.22	0.12	32.7	0.23	0.09	42.8
Reading PSSA, grade 5	0.20	0.11	33.8	0.21	0.08	42.6
Reading PSSA, grade 6	0.16	0.09	31.1	0.18	0.08	41.7
Reading PSSA, grade 7	0.14	0.08	36.8	0.19	0.08	44.2
Reading PSSA, grade 8	0.12	0.07	32.9	0.19	0.08	41.6
Writing PSSA, grade 5	0.37	0.13	53.0	0.36	0.10	61.5
Writing PSSA, grade 8	0.29	0.09	57.5	0.34	0.10	58.9
Science PSSA, grade 4	0.26	0.12	39.9	0.27	0.09	51.9
Science PSSA, grade 8	0.17	0.07	49.0	0.25	0.07	58.5
Combined teacher-level estimates (pre-shrinkage)	0.21	0.09	43.0	0.21	0.06	53.4
Combined teacher-level estimates (post-shrinkage)	0.21	0.08	39.1	0.21	0.06	51.2

Source: Mathematica calculations based on Pennsylvania student data.

Notes: The VAMs are based on statewide samples of teachers and students. Teachers' VAM estimates are based on students in their classrooms at any time during the specified analysis periods.

One *z*-score unit is equal to one standard deviation of student outcomes. One standard deviation of student outcomes is equal to 227 PSSA points in math, 220 points in reading, 276 points in writing, and 190 points in science.

The 84th minus 50th percentile of underlying VAM estimates is the estimated difference in "underlying" value added for the teachers at these percentiles (that is, perfect measures of value added that do not have any estimation error). This is calculated as the standard deviation of value-added estimates with an adjustment for the amount of estimation error using the method in Morris (1983).

All estimates for individual subject-grade combinations are pre-shrinkage.

PSSA = Pennsylvania System of School Assessment; VAM = value-added model.

APPENDIX D

RELATIONSHIPS BETWEEN RUBRIC SCORES AND VALUE ADDED FOR VAMS BASED ON DIFFERENT COHORTS OF STUDENTS

In this appendix, we provide correlations between rubric scores and value added based on value-added models (VAMs) that include different student cohorts. Our main correlational results in Table IV.1 are based on the three-cohort VAM that ends in 2011–2012. In Tables D.1, D.2, and D.3, we report correlations based VAMs that include different cohorts and school years. In Table D.4, we report correlations that are based on a consistent sample of teachers for all four VAMs.

Table D.1. Validity Estimates Based on One-Cohort 2011–2012 Value-Added Model

	All Phase 2		Pittsburgh		Not Pittsburgh	
	Correlation	Teachers	Correlation	Teachers	Correlation	Teachers
Domain 1: Planning and Preparation	0.23*	578	0.21*	329	0.19*	249
Domain 2: Classroom Environment	0.20*	579	0.17*	332	0.18*	247
Domain 3: Instruction	0.29*	601	0.29*	349	0.23*	252
Domain 4: Professional Responsibilities	0.25*	576	0.26*	328	0.17*	248
Professional Practice Rating	0.27*	607	0.26*	355	0.23*	252
1a: Demonstrating knowledge of content and pedagogy	0.11*	411	0.06	235	0.14	176
1b: Demonstrating knowledge of students	0.23*	532	0.21*	328	0.21*	204
1c: Setting instructional outcomes	0.12*	523	0.10	328	0.11	195
1d: Demonstrating knowledge of resources	0.23*	406	0.20*	235	0.21*	171
1e: Planning coherent instruction	0.19*	481	0.18*	236	0.14*	245
1f: Designing ongoing formative assessments	0.19*	411	0.17*	234	0.15	177
2a: Creating a learning environment of respect and rapport	0.16*	454	0.13	244	0.16*	210
2b: Establishing a culture for learning	0.17*	518	0.15*	331	0.18*	187
2c: Managing classroom procedures	0.22*	439	0.26*	241	0.14*	198
2d: Managing student behavior	0.14*	530	0.14*	328	0.09	202
2e: Organizing physical space	0.12*	414	0.09	234	0.11	180
3a: Communicating with students	0.26*	400	0.26*	234	0.20*	166
3b: Using questioning and discussion techniques	0.26*	498	0.27*	332	0.22*	166
3c: Engaging students in learning	0.22*	582	0.19*	333	0.20*	249
3d: Using assessment to inform instruction	0.15*	584	0.17*	336	0.07	248
3e: Demonstrating flexibility and responsiveness	0.21*	396	0.22*	233	0.15	163
4a: Reflecting on teaching and student learning	0.20*	550	0.23*	327	0.12	223
4b: System for managing students' data	0.12*	521	0.05	328	0.18*	193
4c: Communicating with families	0.18*	501	0.21*	327	0.08	174
4d: Participating in a professional community	0.22*	412	0.17*	235	0.22*	177
4e: Growing and developing professionally	0.19*	402	0.15*	231	0.18*	171
4f: Showing professionalism	0.20*	368	0.19*	198	0.15	170

Source: Mathematica calculations based on Phase 2 classroom observation data in the 2011–2012 school year and value-added estimates from 2011–2012.

Note: Correlations are adjusted for estimation error in value-added estimates (Jacob and Lefgren 2008). An unadjusted correlation is 10 percent smaller on average.

* Statistically significant at the 5 percent level.

Table D.2. Validity Estimates Based on Three-Cohort Value-Added Model Ending with 2010–2011

	All Phase 2		Pittsburgh		Not Pittsburgh	
	Correlation	Teachers	Correlation	Teachers	Correlation	Teachers
Domain 1: Planning and Preparation	0.18*	528	0.19*	296	0.18*	232
Domain 2: Classroom Environment	0.17*	529	0.18*	299	0.15*	230
Domain 3: Instruction	0.24*	550	0.26*	315	0.22*	235
Domain 4: Professional Responsibilities	0.07	524	0.08	295	0.06	229
Professional Practice Rating	0.20*	556	0.20*	321	0.19*	235
1a: Demonstrating knowledge of content and pedagogy	0.09	369	0.07	204	0.13	165
1b: Demonstrating knowledge of students	0.13*	483	0.13	295	0.12	188
1c: Setting instructional outcomes	0.14*	474	0.12	295	0.19*	179
1d: Demonstrating knowledge of resources	0.06	365	0.06	206	0.07	159
1e: Planning coherent instruction	0.12*	434	0.16*	207	0.07	227
1f: Designing ongoing formative assessments	0.13*	373	0.16*	206	0.09	167
2a: Creating a learning environment of respect and rapport	0.16*	409	0.16*	213	0.17*	196
2b: Establishing a culture for learning	0.22*	475	0.24*	298	0.20*	177
2c: Managing classroom procedures	0.10*	396	0.13	207	0.06	189
2d: Managing student behavior	0.14*	484	0.15*	295	0.13*	189
2e: Organizing physical space	−0.01	374	−0.03	205	0.02	169
3a: Communicating with students	0.20*	359	0.20*	202	0.21*	157
3b: Using questioning and discussion techniques	0.22*	456	0.25*	298	0.17*	158
3c: Engaging students in learning	0.19*	530	0.23*	299	0.13*	231
3d: Using assessment to inform instruction	0.16*	535	0.18*	303	0.11*	232
3e: Demonstrating flexibility and responsiveness	0.16*	360	0.15*	205	0.19*	155
4a: Reflecting on teaching and student learning	0.07	498	0.07	294	0.08	204
4b: System for managing students' data	0.05	470	0.07	295	−0.02	175
4c: Communicating with families	0.04	457	0.07	294	−0.06	163
4d: Participating in a professional community	0.06	370	0.01	205	0.11	165
4e: Growing and developing professionally	0.00	356	0.00	200	−0.02	156
4f: Showing professionalism	0.04	323	−0.03	167	0.12	156

Source: Mathematica calculations based on Phase 2 classroom observation data in the 2011–2012 school year and value-added estimates using data from 2008–2009 through 2010–2011.

Notes: Correlations are adjusted for estimation error in value-added estimates (Jacob and Lefgren 2008). An unadjusted correlation is 6 percent smaller on average.

Because of data limitations, the three-cohort VAM excluding 2011–2012 is based on three cohorts of student achievement growth data outside Pittsburgh, but is limited to two cohorts—2009–2010 and 2010–2011—for teachers in Pittsburgh.

* Statistically significant at the 5 percent level.

Table D.3. Validity Estimates Based on One-Cohort 2010–2011 Value-Added Model

	All Phase 2		Pittsburgh		Not Pittsburgh	
	Correlation	Teachers	Correlation	Teachers	Correlation	Teachers
Domain 1: Planning and Preparation	0.14*	498	0.15*	272	0.11	226
Domain 2: Classroom Environment	0.14*	500	0.13*	275	0.14*	225
Domain 3: Instruction	0.20*	520	0.17*	291	0.25*	229
Domain 4: Professional Responsibilities	0.09*	496	0.11	271	0.07	225
Professional Practice Rating	0.16*	526	0.14*	297	0.19*	229
1a: Demonstrating knowledge of content and pedagogy	0.12*	351	0.08	188	0.18*	163
1b: Demonstrating knowledge of students	0.10	456	0.09	271	0.09	185
1c: Setting instructional outcomes	0.12*	448	0.12	271	0.13	177
1d: Demonstrating knowledge of resources	0.01	348	0.03	190	−0.02	158
1e: Planning coherent instruction	0.11*	413	0.18*	191	0.05	222
1f: Designing ongoing formative assessments	0.08	353	0.11	190	0.02	163
2a: Creating a learning environment of respect and rapport	0.15*	387	0.13*	195	0.17*	192
2b: Establishing a culture for learning	0.18*	448	0.16*	274	0.23*	174
2c: Managing classroom procedures	0.09	374	0.09	189	0.09	185
2d: Managing student behavior	0.12*	455	0.16*	271	0.05	184
2e: Organizing physical space	0.01	355	−0.01	189	0.04	166
3a: Communicating with students	0.24*	338	0.19*	184	0.31*	154
3b: Using questioning and discussion techniques	0.19*	429	0.16*	274	0.27*	155
3c: Engaging students in learning	0.16*	502	0.14*	276	0.19*	226
3d: Using assessment to inform instruction	0.14*	505	0.15*	279	0.11	226
3e: Demonstrating flexibility and responsiveness	0.14*	340	0.11	189	0.19*	151
4a: Reflecting on teaching and student learning	0.09	470	0.11	270	0.05	200
4b: System for managing students' data	0.08	443	0.08	271	0.06	172
4c: Communicating with families	0.02	430	0.06	270	−0.11	160
4d: Participating in a professional community	0.06	353	0.05	189	0.07	164
4e: Growing and developing professionally	0.04	339	0.06	184	−0.01	155
4f: Showing professionalism	0.08	306	0.00	152	0.17*	154

Source: Mathematica calculations based on Phase 2 classroom observation data in the 2011–2012 school year and value-added estimates from 2010–2011.

Note: Correlations are adjusted for estimation error in value-added estimates (Jacob and Lefgren 2008). An unadjusted correlation is 9 percent smaller on average.

* Statistically significant at the 5 percent level.

Table D.4. Validity of Rubric Domains and Professional Practice Rating by Value-Added Model—All Districts and Consistent Samples of Teachers

	Three-Cohort Value-Added Models				One-Cohort Value-Added Models			
	2009–2010 Through 2011–2012		2008–2009 Through 2010–2011		2011–2012		2010–2011	
	Corr.	Teachers	Corr.	Teachers	Corr.	Teachers	Corr.	Teachers
Domain 1: Planning and Preparation	0.23*	480	0.17*	480	0.23*	480	0.16*	480
Domain 2: Classroom Environment	0.20*	482	0.15*	482	0.19*	482	0.15*	482
Domain 3: Instruction	0.26*	502	0.20*	502	0.29*	502	0.20*	502
Domain 4: Professional Responsibilities	0.19*	478	0.09	478	0.27*	478	0.12*	478
Professional Practice Rating	0.23*	508	0.17*	508	0.27*	508	0.17*	508

Source: Mathematica calculations based on Phase 2 classroom observation data in the 2011–2012 school year and value-added estimates using data from 2008–2009 through 2011–2012.

Notes: Because of data limitations, the three-cohort VAM that excludes 2011–2012 does not include students from the 2008–2009 school year for Pittsburgh teachers.

Correlations are adjusted for estimation error in value-added estimates (Jacob and Lefgren 2008).

* Statistically significant at the 5 percent level.

MATHEMATICA
Policy Research

www.mathematica-mpr.com

Improving public well-being by conducting high-quality, objective research and surveys

Princeton, NJ ■ Ann Arbor, MI ■ Cambridge, MA ■ Chicago, IL ■ Oakland, CA ■ Washington, DC